



# Optimiser ses recherches dans Google

Le 21 juillet 2008, à 12:44 par Ulhume...

Google est une mine d'information, ce n'est un secret pour personne. Et autant il m'arrive de taper sur certaines dérives du géant de la recherche, autant je dois bien avouer que leur moteur est d'une rare puissance. Il est rapide dans ses résultats, écume le web à une vitesse étourdissante (il n'est plus rare de voir un contenu indexé dans les heures qui suivent sa mise en ligne), et dispose d'un langage de formulation des demandes aussi vaste que méconnu. Et c'est cet aspect paradoxalement un peu obscure que je vais essayer de couvrir dans ce tutoriel : la recherche de documents sur le WEB, et seulement ça...



Certaines techniques de recherche permettent de mettre à jour la bêtise de nombre de sociétés, d'organismes ou de particuliers qui mettent en ligne, souvent sans le savoir, un volume impressionnant de données. Je ne peux donc être tenu responsable de ce que vous ferrez des résultats de ces recherches.

De mon point de vue, ce tutoriel s'arrête à la page de résultat de Google. Ce que vous faites des liens qui s'y trouvent est de **vos** responsabilité.

Je vous rappelle seulement, à toutes fins utiles, que lorsque vous téléchargez un fichier protégé **à ne pas diffuser**, qui sont paradoxalement disponibles à profusion sur le net, je n'ai aucune idée du risque encouru à les parcourir. J'imagine que cela doit varier si ces derniers proviennent de la petite PME du coin, ou du ministère de la défense... Si un juriste a des informations à ce sujet, je suis preneur.

En tout cas pour moi c'est ceinture/bretelle, vous savez ce que vous faites.

## De la question à la requête

De temps en temps, lorsque je n'ai rien de plus intelligent à faire, je jette un oeil aux statistiques de mes sites (qui n'utilisent pas de [scripts externes](#) <sup>[1]</sup> ;p ) et plus particulièrement à la très instructive liste de phrases qui, données à manger aux automates de recherche, aboutissent chez moi. Et il n'est pas rare, de tomber sur des formulations d'un académisme aussi irréprochable qu'inefficace. Comme par exemple :

[Comment rendre mon Windows XP légal ?](#) <sup>[2]</sup>



Petite note pratique: Toutes les requêtes données en exemples sont cliquables, n'hésitez pas à les essayer...

Alors, il y a plusieurs choses remarquables concernant la personne qui a soumis cette phrase à Google :

- Elle pose **une question** à M. Google et met en conséquence toutes les ponctuations d'usage.
- Elle spécifie bien qu'il s'agit de **Son** Windows et pas celui du voisin.
- Enfin, elle met bien la majuscule en début de phrase et sur le nom-propre.

La personne qui a écrit cela doit j'imagine s'attendre à une réponse du type **Pour rendre Ton Windows XP légal, cher ami, il faudrait tout d'abord songer à l'acheter...** Malheureusement, M. Google n'est pas aussi humain et ne comprend rien à la grammaire et encore moins au sens des mots. Voyons donc un peu ce que Google comprend vraiment...

## De la question à la liste de mots

En prenant l'exemple de cet utilisateur qui a tapé **Comment rendre mon Windows XP légal ?**, il ne s'agit pas de se moquer mais bien au contraire de comprendre ce que Google fait dans la vie. Son métier est de lire le web de A à Z. Chaque lien, chaque page, chaque image est lu et stocké dans sa grande base. Ceci fait, il va analyser chaque **mot** de chaque **page** et alimenter ainsi d'énormes **indexes**, comme celui d'une bibliothèque. Il sait donc qu'il connaît très exactement **30 700 000** pages qui contiennent le mot **fenêtre**. Il ne sait en revanche rien du sens de ce mot, et encore moins du contexte sémantique de son utilisation. Une **publicité pour remplacer les fenêtres d'un pavillon** ou un **tutoriel traitant de la manipulation des fenêtres dans un ordinateur** parlent pour lui de la même chose... Ainsi les ponctuations, les mots comme **Comment, Pourquoi, etc..** peuvent être sans trop de risque éliminés.

Dans la même idée, Google ne fait aucune différence entre majuscules et minuscules. Mieux, il ne comprends même pas les accents et le mot **été** et **ete** sont pour lui identiques.

Et même les mots eux-mêmes ne sont pas tous logés à la même enseigne. A chaque mots d'une phrase, Google va associer un **poind**. Et les "petits" mots, typiquement les pronoms (un, le, mon, etc.) sont pour lui des mots faibles, presque inexistant. Autant ne même pas les taper tant ils ne servent à rien.

Fort de tout cela, notre phrase d'introduction peut donc, avec le même résultat, être reformulé de la manière suivante :

[rendre windows xp legal](#) <sup>[3]</sup>

Tout simplement...

## Les expressions avancées

Pour l'instant, avec cette formulation, google cherche individuellement chacun des mots que nous lui avons donnés. Or dans cet exemple, c'est bien **Windows XP** qui est au centre de notre recherche. Nous n'avons pas envie d'avoir, même dans une seule page, un **Windows** d'un côté et un **XP** trente kilomètres plus loin... Heureusement, nous avons la possibilité de **grouper** ces deux mots en une seule expression. Ce groupement s'écrit en plaçant les mots entre double-guillemets.

Une formulation plus efficace sera donc :

[rendre "windows xp" legal](#) <sup>[4]</sup>

Pour améliorer encore notre demande, nous avons aussi la possibilité de demander à Google de garder nos expressions (mots ou groupes de mot) les plus proches possibles les unes des autres grâce au symbole **\*** . En langage Google ce symbole signifie en gros **un ou plusieurs mots, peu importe lesquels**.

Nous obtenons ainsi la version finale de notre demande à Google qui va donc permettre d'obtenir la liste des pages contenant les expressions **rendre windows XP légal** les plus proches possibles les unes des

autres :

[rendre \\* "windows xp" \\* legal](#) <sup>[5]</sup>

## Conclusion

En allant de la formulation d'origine **Comment rendre mon Windows XP légal ?** à la phrase finale **rendre \* "windows xp" \* legal**, nous comprenons bien que l'on ne pose plus une **question** à Google mais que nous lui formulons une **requête** sous la forme d'une **liste de mot ou d'expressions** à rechercher. Une requête pour l'instant très simple mais que nous allons rapidement pousser beaucoup plus loin.



### Notes complémentaires:

1. Google se fiche de l'ordre des mots pour ses résultats. En revanche, l'ordre des mots va influencer l'ordre dans lequel les résultats sont classés en donnant ainsi plus de "force" à un premier mot, un peu moins au second, etc... S'il n'y a que 10 résultats cela n'a pas beaucoup d'importance, mais lorsqu'il y en a des milliers, cela peut changer la donne ;-). C'est ainsi que les "petits" mots, ils ne seront pas vraiment ignorés, mais pondéré plus faiblement.
2. Google ne comprend pas les accents, mais comprends les ligatures. Ainsi le mot `b?ufs` donnera le même résultat que `boeuf`.
3. A chaque requête, Google affiche le nombre de résultat total trouvé, dans la petite zone bleue qui se trouve entre la zone de saisie et le résultat. Dans notre exemple, il y a 708 000 réponses et Google affiche les 10 premières. Si vous allez dans les préférences (à droite de la zone de saisie), vous pouvez afficher non pas 10, mais 20, 30, 50 ou 100 réponses par page. Mais cela sera un peu plus long à arriver.
4. A noter aussi que le nombre maximal de membre dans une requête Google est de 10. Les autres seront ignorés.
5. Mis à part un mot, et une expression, Google connaît encore un autre type de chose qu'il peut chercher, les intervalles de nombre. Je ne sais pas bien à quoi cela peut servir mais ça existe et vous saurez peut-être en faire quelque chose. Ainsi si vous tapez `Présidentielles 1980..2020`, vous aurez toutes les pages contenant **présidentielles 2000**, **présidentielles 2001**, etc.. jusqu'à **présidentielles 2020**.
6. Autre aspect intéressant, Google est meilleur en orthographe que vous (en tout cas clairement que moi). Ainsi si un mot est mal orthographié, il vous proposera de corriger votre requête avec la bonne syntaxe.

## Les opérateurs logiques

### Des ET et des OU...

Comme nous l'avons vu, une requête Google est composée d' **expressions** à chercher (ex. legal ou "windows xp"). Nous allons voir maintenant comment ces expressions sont liées entre elles.

Par défaut, Google cherche **toutes** les expressions que vous saisissez sur une même page. Il place

implicitement des **ET** invisibles entre chacun d'entre eux. En anglais **ET** s'écrivant **AND**, notre requête peut aussi s'écrire de la manière suivante :

[rendre AND windows AND legal](#) [6]

Ou encore (AND et & étant synonymes)

[rendre & "windows xp" & legal](#) [7]

**AND** est ce que l'on appelle un **opérateur logique**. Un opérateur logique va lier... logiquement une expression à une autre. L'autre opérateur logique évident est le **OU**. Par exemple, si nous voulons améliorer la recherche précédente et chercher les documents qui contiennent **légal** OU **valide**, nous formulerions notre requête ainsi (OU s'écrit "OR" en anglais) :

[rendre AND "windows xp" AND legal OR valide](#) [8]

Et comme les AND peuvent être sous-entendus, cela nous donne :

[rendre "windows xp" legal OR valide](#) [9]

Pour faire encore plus compacte, le OR qui peut aussi s'écrire par le symbole **|**. Ce qui nous donne la requête :

[rendre "windows xp" legal | valide](#) [10]

Enfin, il est possible de grouper les opérateurs avec des parenthèses pour lever les ambiguïtés, comme par exemple :

[rendre "window xp" \(legal | valide | cool\)](#) [11]

## L'exclusion

Nous savons maintenant comment indiquer à Google ce que nous voulons dans les résultats. Essayons maintenant de lui signifier ce que nous ne voulons pas. En logique c'est ce que l'appelle l'opérateur **NON** qui s'écrit en langage Google par le signe **-** suivi de l'expression (mot ou group de mot) à exclure. Donc si par exemple, je ne veux pas qu'apparaisse le mot **téléchargement**, cela donne :

[rendre "window xp" \(legal | valide | cool\) -téléchargement](#) [12]

## Les modificateurs

Nous savons déjà construire des requêtes complexes, avec des opérateurs logiques, des expressions et des termes exclus. Nous allons maintenant aborder la partie la plus intéressante du système de recherche de Google, **les modificateurs**.

De manière générale, les modificateurs permettent d'altérer la manière dont une expression est prise en compte dans la recherche. Cela peut changer son importance, inclure ses synonymes, et spécifier précisément où l'expression doit être trouvée.

## La modification de l'importance d'une expression

Google place des ET invisibles dans la requête. Ainsi un mot trop court, ou trop peu significatif par rapport aux autres, risque d'être zappé de la recherche. La plupart du temps c'est un comportement pratique , mais il arrive que l'on **tienne** à un mot, même si Google le trouve sans intérêt. C'est là qu'intervient le signe `+` suivi de l'expression à forcer.

Si par exemple nous désirons forcer le mot `mon` dans la requête précédente, cela donnerait :

[rendre +mon windows legal](#) [13]

## Les synonymes

Google peut vous permettre de "flouter" un peu vos recherches en lui demandant de chercher un mot ET ses synonymes. Parler de synonymes n'est pas totalement exact. Il s'agit plutôt d'une relation statistique généralement constatée par Google entre des mots. Par exemple si vous tapez :

[~bateau](#) [14]

Vous aurez aussi des résultats concernant les.. croisières !! De l'influence du web marchand sur Google ;-)  
)

## Spécifier où chercher

D'abord une page web c'est quoi exactement ? Et bien c'est tout d'abord une adresse (url). Puis un titre (title) et un contenu. Enfin ce sont des liens composés chacun d'un texte (anchor) et d'une adresse de destination (link).

Toutes les requêtes que nous avons vues précédemment portaient exclusivement sur le **contenu** de la page (ou du document). Mais il est possible de demander à Google que certaines expressions soient recherchées dans une des parties spécifiques de la page comme le titre, l'url, un lien, ou le texte d'un lien.

Pour prendre un exemple, imaginons que nous voulions rechercher toutes les pages qui contiennent le mot **drm** et dont le **titre** contient le mot `DADVSI`. Nous formulerons alors la requête suivante :

[drm intitle:dadvs](#) [15]

La première partie est classique, elle indique que **drm** doit être cherché dans le contenu de la page. La seconde expression en revanche, utilise le modificateur `intitle:` qui restreint la recherche du mot **dadvs** aux seuls titres (title). Et si nous voulons trouver non pas un mot simple mais un groupe de mots dans ce titre, nous allons utiliser le modificateur `allintitle:` :

[drm allintitle:"loi dadvs"](#) [16]

De la même manière si nous voulons chercher dans le texte d'un lien, nous utiliserons le modificateur `inanchor:`, ou `allinanchor:` s'il s'agit d'un groupe de mots.

[allanchor:dadvs](#) [17]

[allinanchor:"dadvs drm"](#) [18]

Pour chercher un mot ou un groupe de mots dans l'**adresse** même des pages, nous utiliserons les modificateurs `inurl:` et `allinurl:`. Ainsi, pour partir à la chasse aux étourdis, nous pouvons tester ce modificateur avec cette requête :

[inurl:userfiles](#) [19]

[allinurl:"userfiles media"](#) [20]

Enfin, si nous voulons chercher les pages qui contiennent au moins un lien qui pointe vers une autre page nous utiliserons le modificateur `link:`. Par exemple, si nous voulons connaître tous les sites qui référencent la page `http://moutons.karma-lab.net/node/10`, nous écrivons la requête suivante :

[link:moutons.karma-lab.net/node/10](#) [21]

Bien évidemment, tous ces modificateurs peuvent être combinés par des liens logiques. Par exemple, au hasard balthazar, essayons ceci :

[intitle:"index of" \( link:mp3 | link:ogg\)](#) [22]

Un grand classique qui je l'espère vous est maintenant facile à comprendre et à décortiquer ;-)

Un autre exemple amusant est de rechercher ces petites caméras connectées à internet et que beaucoup semblent oublier de protéger. Or si l'on sait ce que contient l'écran de contrôle d'une de ces caméras, il devient alors très simple d'en obtenir la liste :

[inurl:"MultiCameraFrame?Mode=" OR inurl:"ViewerFrame?Mode=" OR inurl:"/view/view.shtml?videos="](#)

[23]

C'est comme cela, en faisant gogoter une caméra collée sous leur nez, que je me suis amusé 10 bonnes minutes à affoler les girafes dans un Zoo quelque part sur la planète ;-)

## Les expressions spéciales

### Le type de document

Google indexe à peu près tout ce qu'il trouve et même s'il était à l'origine limité aux pages web, son action c'est vite étendue aux documents de traitement de texte, aux feuilles de calcul, etc. A ma connaissance, les types de document indexés sont les suivants.

Pour les documents texte, les formats reconnus sont :

- Simples (ANS, TXT)
- OpenDocument (ODT)
- Portable Document File (PDF)
- PostScript (PS)
- Lotus WordPro (LWP)
- Microsoft Word (DOC)
- Microsoft Write (WRI)
- Rich Text Format (RTF)
- MacWrite (MW)

Pour les feuilles de calcul :

- OpenCalc (ODS)
- Lotus 1-2-3 (WK1, WK2, WK3, WK4, WK5, WKI, WKS, WKU)
- Microsoft Excel (XLS)

### Les présentations :

- OpenPresentation (ODP)
- Microsoft PowerPoint (PPT)

### Et divers autres :

- Shockwave (SWF)
- Autodesk (DWF)
- Google Earth (KLM,KMZ)
- Microsoft Works (WDB, WKS, WPS)

Sachant cela, il est possible d'ajouter à notre recherche un ou plusieurs types de document désiré. Par exemple, imaginons que nous voulions connaître les documents PDF ou DOC que des étourdis ont mis en ligne et qui ont le malheur d'être confidentiels :

[+"document provisoire" +"ne pas diffuser" filetype:doc | filetype:pdf](#) [24]

C'est à ce stade que l'on se rend compte du travail qu'il y a encore à réaliser en terme d'éducation..Car google lui, ne pardonne pas...

De la même manière il est possible d'enlever des types de documents à une recherche en utilisant l'opérateur d'exclusion. Ainsi l'exemple suivant permet de rechercher tous les documents qui ont comme titre **Spécifications techniques** en excluant les fichiers .pdf :

[allintitle:"specifications techniques" -filetype:pdf](#) [25]

## Le domaine d'une page

Google sait évidemment d'où viennent les documents qu'il indexe. Et il est possible grâce à l'expression `site:` d'utiliser cette information dans ses recherches. Par exemple si nous voulons connaître la liste des documents PDF mis en ligne via le site `unesco.org` OU `europa.eu` , nous fabriquerons la requêtes suivante :

[+filetype:pdf site:unesco.org | site:europa.eu](#) [26]

Le modificateur **site:** permet de recherche tous les noms de sites se terminant par l'expression qu'il désigne (ici `unesco.org` ou `europa.eu`). En étant plus précis, il est donc possible de limiter encore la recherche au seul documents du site `unesdoc.unesco.org` :

[+filetype:pdf site:unesdoc.unesco.org](#) [27]

Dans la même idée, il est possible non pas d'inclure mais d'éliminer toutes les pages venant d' un domaine donnée:

[rendre \\* "windows xp" \\* legal -site:microsoft.fr -site:microsoft.com](#) [28]

## Les sous-moteurs

Récemment je suis tombé sur un sous-moteur dédié à Linux. Après quelques recherches, ils sont finalement plus nombreux que je l'imaginai. Alors en plus de ce que vous avez classiquement (texte, images, news, etc..), nous avons aussi :

- [Les recherches pour Linux](#) [29], par exemple [Atheros](#) [30].
- [Les recherches pour \\*BSD](#) [31], par exemple [Atheros](#) [32].
- [les recherches pour MacOS](#) [33], par exemple [Atheros](#) [34].
- [les recherches pour Microsoft](#) [35], par exemple [Atheros](#) [36] (Rho0, le premier lien est un plantage ;-)).

Il existe aussi des sous-moteurs dédiés à des fonds documentaires particuliers :

- [Les recherches sur les brevets](#) [37]. Il est possible de chercher les brevets en texte simple ou par l'utilisation de modificateurs comme `patent:` pour trouver par numéro. Par exemple, le célèbre brevet de British Telecom sur [l'hyperlien](#) [38].
- Pour les Fox Mulder en herbe, [les recherches sur les documents gouvernementaux américains](#) [39]. Par exemple [Roswell UFO](#) [40].
- Enfin, pour rechercher exclusivement dans les [articles de journaux](#) [41], par exemple [ceci](#) [42]. Donc ça marche très bien ;-).
- [Recherches de livres](#) [43], par exemple [Guerre et Paix](#) [44].
- [Recherche parmi les blogs](#) [45]. Mais il n'est absolument pas pertinent !! La preuve avec [cet exemple](#) [46] alors qu'il est clairement écrit en slogan de ce site, que ce n'est pas un blog !!!.

## Conclusion

Voilà, fin du petit tour d'horizon de ce qu'il est possible de faire comme recherche sur le WEB avec Google. Comme vous l'avez vu, les possibilités sont nombreuses et juste limité, comme beaucoup de choses en ce domaine, par l'imagination.

### Liens:

- [1] <http://artisan.karma-lab.net/comment-ne-plus-etre-trace-par-google-analytics>
- [2] <http://www.Google.fr/search?q=Comment rendre mon Windows XP légal ?>
- [3] <http://www.Google.fr/search?q=rendre windows xp legal>
- [4] [http://www.Google.fr/search?q=rendre "windows xp" legal](http://www.Google.fr/search?q=rendre )
- [5] [http://www.Google.fr/search?q=rendre \\* "windows xp" \\* legal](http://www.Google.fr/search?q=rendre * )
- [6] <http://www.Google.fr/search?q=rendre AND windows AND legal>
- [7] [http://www.Google.fr/search?q=rendre & "windows xp" & legal](http://www.Google.fr/search?q=rendre & )
- [8] [http://www.Google.fr/search?q=rendre AND "windows xp" AND legal OR valide](http://www.Google.fr/search?q=rendre AND )
- [9] [http://www.Google.fr/search?q=rendre "windows xp" legal OR valide](http://www.Google.fr/search?q=rendre )
- [10] [http://www.Google.fr/search?q=rendre "windows xp" legal | valide](http://www.Google.fr/search?q=rendre )
- [11] [http://www.Google.fr/search?q=rendre "window xp" \(legal | valide | cool\)](http://www.Google.fr/search?q=rendre )
- [12] [http://www.Google.fr/search?q=rendre "window xp" \(legal | valide | cool\) -téléchargement](http://www.Google.fr/search?q=rendre )
- [13] <http://www.Google.fr/search?q=rendre +mon windows legal>
- [14] <http://www.Google.fr/search?q=~bateau>
- [15] <http://www.Google.fr/search?q=drm intitle:dadvsi>
- [16] [http://www.Google.fr/search?q=drm allintitle:"loi dadvsi"](http://www.Google.fr/search?q=drm allintitle:)
- [17] <http://www.Google.fr/search?q=allanchor:dadvsi>
- [18] [http://www.Google.fr/search?q=allinanchor:"dadvsi drm"](http://www.Google.fr/search?q=allinanchor:)
- [19] <http://www.Google.fr/search?q=inurl:userfiles>
- [20] [http://www.Google.fr/search?q=allinurl:"userfiles media"](http://www.Google.fr/search?q=allinurl:)
- [21] <http://www.Google.fr/search?q=link:moutons.karma-lab.net/node/10>

- [22] [http://www.Google.fr/search?q=intitle:"index of" \( link:mp3 | link:ogg\)](http://www.Google.fr/search?q=intitle:)
- [23] [http://www.Google.fr/search?q=inurl:"MultiCameraFrame?Mode=" OR inurl:"ViewerFrame?Mode=" OR inurl:"/view/view.shtml?videos="](http://www.Google.fr/search?q=inurl:)
- [24] [http://www.Google.fr/search?q="+document provisoire" +"ne pas diffuser" filetype:doc | filetype:pdf](http://www.Google.fr/search?q=)
- [25] [http://www.Google.fr/search?q=allintitle:"specifications techniques" -filetype:pdf](http://www.Google.fr/search?q=allintitle:)
- [26] [http://www.Google.fr/search?q="+filetype:pdf site:unesco.org | site:europa.eu](http://www.Google.fr/search?q=)
- [27] [http://www.Google.fr/search?q="+filetype:pdf site:unesdoc.unesco.org](http://www.Google.fr/search?q=)
- [28] [http://www.Google.fr/search?q=rendre \\* "windows xp" \\* legal -site:microsoft.fr -site:microsoft.com](http://www.Google.fr/search?q=rendre * )
- [29] <http://www.google.com/linux>
- [30] <http://www.google.com/linux?hl=fr&q=atheros&btnG=Rechercher&lr=>
- [31] <http://www.google.com/bsd>
- [32] <http://www.google.com/bsd?hl=fr&q=atheros&btnG=Rechercher&lr=>
- [33] <http://www.google.com/mac.html>
- [34] <http://www.google.com/mac?hl=fr&q=atheros&btnG=Rechercher&lr=>
- [35] <http://www.google.com/microsoft.html>
- [36] <http://www.google.com/microsoft?hl=fr&q=atheros&btnG=Rechercher&lr=>
- [37] <http://www.google.com/patents>
- [38] <http://www.google.com/patents?q=patent:4873662>
- [39] <http://www.google.com/unclesam>
- [40] <http://www.google.com/search?site=unclesam&restrict=unclesam&hl=fr&output=unclesam&q=roswell ufo>
- [41] <http://news.google.com/archivesearch>
- [42] <http://news.google.com/archivesearch?q=nabot président>
- [43] <http://www.google.com/books>
- [44] <http://www.google.com/books?q=guerre et paix&hl=fr&spell=1&oi=spell>
- [45] <http://blogsearch.google.com/>
- [46] <http://blogsearch.google.com/blogsearch?hl=fr&q=artisan>