Recherches sur Internet: méthode et astuces

Version	*Date*	*Contributeur*
1.1	*Septembre 2006*	Mise à jour selon l'évolution des moteurs de recherche
1.0	*Mai 2006*	Christophe Dupriez, dupriez@destin.be

Ce texte est distribué sous licence Creative Commons 2.0 dans sa variante "Paternité à maintenir – Usage non commercial – Partage sous licence identique":
http://creativecommons.org/licenses/by-nc-sa/2.0/be/legalcode.fr

Table des matières

Problématique d'ensemble....1
Si vous avez autre chose à faire….....1
Comment font les autres ?....1
Du questionnement aux réponses, tout un voyage!....2
Qu'est ce qu'un moteur de recherche sur Internet....2
Des idées aux mots….....3
Des mots aux idées….....4
Choisir de bons mots-clés ?....4
1. Le niveau sémantique....5
2. Le niveau terminologique....5
3. Le niveau lexical....5
Moteurs de recherche....6
Le moteur de recherche Google....6
Termes à chercher:....6
Opérateurs logiques (booléens):....7
Limites:....7
Dates:....7
Tri du résultat:....8
Le moteur de recherche Exalead.com....10
Termes à chercher:....10
Opérateurs logiques (booléens):....11
Tris des résultats:....11
Stratégie de recherche documentaire....12

Procédure de recherche suggérée par l'Université de Berkeley: 12
Procédure complète -- Besoins et solutions 13
Procédure complète -- Pour chaque expression de recherche 13
Petit rappel des opérations booléennes 14
Rappel et précision 14
Grandes étapes dans la préparation d'une recherche Google ou Exalead 15
Adapter sa recherche au type d'information 15
Un objet "identifié": 15
Un "nom propre" 16
Un lieu: 16
Une institution ou une entreprise: 17
Une personne: 17
Un "nom commun" 17
Un objet 18
Un concept 18
Une source 18
Un nombre 19
L'analyse du résultat des recherches 19
Le document: 19
Le contexte d'un document: 19
Les éléments d'un document: 20
Les types de document: 20
Bibliographie sur les méthodes de recherche 21

Ce document sera régulièrement mis à jour: veuillez contacter christophe.dupriez@destin.be pour vous inscrire aux mises à jour, pour signaler des erreurs ou pour faire des suggestions…

Problématique d'ensemble

Si vous avez autre chose à faire…

Si vous n'avez pas le temps ou le courage de faire par vous-même une recherche, telle que ce qui suit, il y a des spécialistes de l'information, des bibliothécaires qui peuvent vous aider:

- Le Guichet du Savoir vous répond en 3 jours ouvrables maximum: http://www.guichetdusavoir.org/GdS/
- Les bibliothèques publiques francophones de Belgique ont uni leurs forces: http://www.bibliothequevirtuelle.be/
- On trouve ce genre de service dans d'autres pays comme les Pays Bas ou l'Angleterre: http://www.questionpoint.org/crs/servlet/org.oclc.home.TFSRedirect?SS_COMMAND=CUST_SUP&Category=COE
- Des encyclopédies peuvent apporter une bonne réponse générale sur un sujet donné. Il faut toutefois garder son esprit critique car la qualité des rubriques est inégale: http://www.wikipedia.fr, http://www.answers.com

Comment font les autres ?

En avril 2006, Harvest Digital a étudié la façon dont 205 internautes britanniques (utilisateurs de Internet depuis 3 ans et y passant plus de 10 heures par semaine) utilisaient les moteurs de recherche:

- Google est utilisé par 94% des internautes mais 76% des internautes utilisent plusieurs moteurs de recherche.
- 47% des personnes interrogées passent plus de 3 heures par semaine rien qu'en recherches
- 68% des internautes utilisent 3 mots clés ou plus pour exprimer leurs recherches
- Au niveau de leurs échecs de recherche, 36% des internautes les imputent à des mots clés qu'ils choisiraient mal. 32% estiment que l'information qu'ils recherchent est trop spécialisée. Seulement 8% pensent que cela pourrait être dû au moteur de recherche.
- 24% des internautes n'apprécient pas la présence de liens sponsorisés.
- L'étude a également demandé ce qui, d'après les internautes, pourrait améliorer leurs résultats : plus d'entraînement et d'expérience pour 50% des sondés, l'utilisation de plusieurs moteurs pour 9% et de meilleurs moteurs pour seulement 5%...

http://www.harvestdigital.com/fact_sheets.cfm

Nous allons essayer d'apporter une réponse à cette moitié des internautes qui demandent à améliorer leur compétence en recherche. Et le choix des mots-clés est évidemment le principal point critique…

Du questionnement aux réponses, tout un voyage!

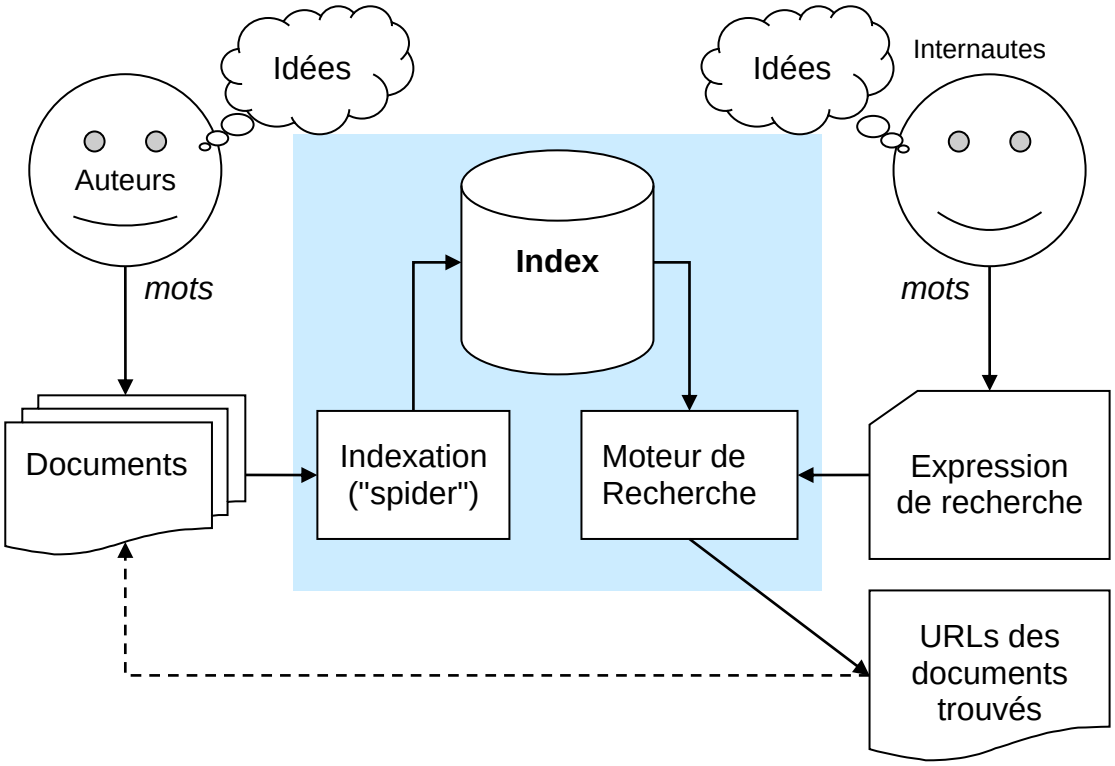
Il y a tant de problèmes et tant de solutions, que le plus difficile est de se rappeler de ceux qui nous importent vraiment !

S'organiser pour ne pas se perdre dans notre voyage sur Internet: comme le Petit Poucet, garder des traces pour se rappeler de ses choix: ***Noter un mot ou l'autre, faire un schéma…*** *Si on est interrompu, si on a suivi un chemin de traverse, si on revient quelques jours plus tard, les pages tracées de notre main ont souvent le pouvoir de nous ramener là où nous en étions dans notre réflexion.*

Des logiciels comme **Scrapbook** ou **NetSnippets** apportent aussi une solution à celui qui veut organiser rapidement les informations recueillies et être en mesure de les republier facilement vers ses collègues: http://amb.vis.ne.jp/mozilla/scrapbook/, http://www.netsnippets.com

Des services Internet comme http://del.icio.us permettent aussi de consigner ses résultats de recherche en les partageant avec d'autres Internautes.

Qu'est ce qu'un moteur de recherche sur Internet



Grâce à Internet, des millions d'auteurs rendent accessibles à tous des milliards de documents.

Des dizaines de "spiders" (ou « web crawlers ») parcourent inlassablement le Web, obtiennent les documents un à un et créent leur index (pour chaque mot apparaissant dans l'un ou l'autre document, quels sont les différents documents qui le contiennent ?). Certains « spiders » gardent une copie du document pour pouvoir le présenter même si l'original disparaît ou pour pouvoir analyser ce qui a changé entre deux passages.

Sur base du travail de leur « spider », les moteurs de recherche permettent à des centaines de millions d'internautes d'exploiter les index pour trouver les documents dont ils ont besoin.

Google exploiterait près de deux cent milles ordinateurs (mars 2006).

Des idées aux mots…

- L'auteur a des idées: il les transcrit par des enchaînements de mots (avec parfois aussi des images et même des sons), dans un ou plusieurs documents inter-reliés.

  Ces documents sont relativement statiques: ils contiennent, en quelque sorte, des réponses préparées à l'avance.

  L'information qui est retrouvée par les moteurs de recherche, c'est celle que le programme d'indexation ("spider") peut trouver en suivant les liens entre documents. En conséquence:

  - ce qui n'est pas écrit n'est pas indexé,
  - ce qui n'est pas déposé dans un serveur accessible de l'Internet n'est pas indexé,
  - ce qui n'est pas lié au document « racine » d'un serveur, en un nombre limité d'étapes (ou directement par un document extérieur) n'est pas indexé,
  - ce qui n'est pas accessible gratuitement n'est pas indexé.

  Et tout ce qui n'est pas indexé n'est évidemment jamais trouvé par les moteurs de recherche…

- Quand on parle avec un spécialiste, celui-ci élabore une réponse en fonction des questions qui lui ont été posées. De la même manière, il y a des applications informatiques qui produisent **dynamiquement** des informations selon les données d'un problème qu'on leur soumet par formulaire.

  C'est le Web "invisible", la partie du Web que les "spiders" ne peuvent pas indexer puisqu'ils ne connaissent pas les données des problèmes !

  Des catalogues ont été dressés par différentes institutions pour trouver ces banques de données invisibles pour les "spiders".

  - Dadi est un répertoire des banques de données gratuites: http://dadi.enssib.fr/
  - GoshMe est un très bon outil pour chercher dans plusieurs banques de données "invisibles" et pour proposer celles qui semblent les plus pertinentes pour un ensemble de mots cherchés: http://www.goshme.com/

- L'internaute a des besoins mais ce qu'il cherche ce sont des solutions: **quels sont les mots que les auteurs ont pu utiliser pour décrire des solutions aux besoins de l'internaute ?**

- Entre les idées de l'auteur et les besoins de l'internaute, il y a:
    - Les mots et la langue de l'auteur
    - Les hypothèses, les axes de solution à ses besoins que l'internaute est capable d'imaginer
    - La langue et les mots de l'internaute

  Comment gérer cette fracture entre les auteurs et les internautes ?

Des mots aux idées…

Heureusement, l'internaute est dans un processus dynamique. Petit à petit, l'internaute peut améliorer la rédaction de ses requêtes de recherche en effectuant les étapes suivantes:

1. Rédiger une **expression de son besoin** (Quoi? Pour quoi? Qui? Pour qui? Comment? Où? Quand?)
2. Rassembler quelques documents qui parlent de son besoin et qui évoquent des axes de solution et s'imprégner des principaux concepts du domaine
3. Rédiger une **expression pour chaque axe de solution** possible
4. Pour **chaque langue** que l'on comprend, indépendamment, choisir de **bons mots clés** (*la recherche terminologique*)
5. Identifier des sources adéquates: **auteurs, institutions, entreprises, banques de données ou sites spécialistes** du problème à résoudre (*navigation "horizontale" ; si nécessaire, chercher à atteindre le Web invisible en utilisant des moteurs de recherche spécialisés*)
6. Trouver des **documents qui apportent l'un ou l'autre élément de réponse** au besoin (*la stratégie de recherche documentaire*)

Une recherche sur Internet, c'est donc un tout un processus dont on a intérêt à conserver les différents éléments dans un dossier.

Nous ne connaissons pas de logiciel qui appuie spécifiquement l'enchaînement de ces étapes mais **Scrapbook**, **NetSnippets** ou **Del.icio.us** sont une base qu'il faudrait compléter avec des outils terminologiques.

On ajuste évidemment la rigueur dans le suivi de cette méthode selon l'importance et la difficulté de sa recherche.

Choisir de bons mots-clés ?

C'est la clé ! Mais on se confronte à trois niveaux de problèmes dont on doit être profondément conscient pour pouvoir les surmonter:

1. Le niveau sémantique

L'information permet les décisions et les décisions permettent l'action.
La valeur des informations est celle des actions qu'elles déclenchent.

C'est dans ce cadre, qui part de la **volonté d'action**, qui passe par la **prise de décisions**, que se trouve la motivation de vos recherches sur Internet.

Le niveau sémantique est donc le **choix des concepts** et surtout des combinaisons de concepts à l'intersection (ET / AND) desquels se trouvent les informations désirées.

La stratégie de recherche documentaire expliquée dans ce document suit cette approche.

C'est assez paradoxal mais, souvent, on ne sait pas vraiment ce qu'on cherche !
On ressent un besoin, on pressent des solutions et c'est dans la confrontation avec ce qui existe (avec ce qu'on voit sur Internet) que les choses se précisent.
La « promenade » est parfois plus féconde que la « recherche »…

2. Le niveau terminologique

Pour chaque concept, le **choix des termes** (un terme étant formé d'un ou de plusieurs mots) est ensuite critique comme expliqué un peu plus loin. Il faut essayer de ne pas oublier des termes possibles pour le concept que l'on désire trouver: on voudra alors trouver des variantes orthographiques, des synonymes, des traductions dans d'autres langues que l'on comprend. On s'aidera de glossaires, de dictionnaires, de textes explicatifs, etc.

On trouve assez facilement des documents avec les termes que l'on utilise soi-même. Le défi est de trouver ceux avec les termes que d'autres utilisent pour le même concept.

3. Le niveau lexical

Ce niveau est le plus technique et peut jouer de mauvais tours car les moteurs de recherche ne donnent pas tous les outils nécessaires pour les problèmes que l'on peut rencontrer à ce niveau. Mentionnons:

- la frontière entre les mots: où commencent-ils, où finissent-ils ? *pipe-line* ou *pipeline* ? *H2O* ou *H 2 O* ?
- l'allemand et le néerlandais permettent de réunir plusieurs mots en un seul
- les alphabets différents d'une langue à une autre: un même nom propre peut être orthographié différemment dans l'alphabet arabe, cyrillique ou japonais
- les accents: "The" vs "thé", "de" vs "dé", "poisson sale" vs "poisson salé", etc.

La ponctuation qui n'a pas d'importance SAUF dans les nombres (ponctuation différente entre l'Amérique et le Système International), dans les formules chimiques, en musique, etc.

Moteurs de recherche

Le moteur de recherche Google

Termes à chercher:

La documentation de Google sur son interprétation des requêtes est pour le moins spartiate. L'évolution du fonctionnement observé montre que ceci est sans doute à dessein pour garder une liberté de changement maximale.
Ce qui suit a d'ailleurs dû être remanié à cause de modifications récentes.

- *H2O* est cherché comme un seul mot et Google ne trouve alors pas les documents avec *H 2 O* ou H_2O dans leur texte. Ceux-ci sont retrouvés en demandant "H 2 O". En théorie H-2-O (voir plus loin le rôle du tiret) devrait trouver aussi bien *H2O* que *H 2 O* et H_2O. Malheureusement, l'opérateur « tiret » ne fonctionne que pour deux mots (par exemple pipe-line) et pas pour trois et plus.
- **mot** : Un mot et ses variantes singulier/pluriel - masculin/féminin – avec/sans accents. Par exemple chevaux d'arçon retrouve *cheval d'arçon*: cet algorithme fonctionne en français et en anglais mais pas en néerlandais (il ne connaît pas les pluriels en "**en**"). Attention : la variante que vous spécifiez est favorisée dans le tri des documents présentés.
- **~mot** : Un mot et ses synonymes. Fonctionne avec un dictionnaire anglais même sur les recherches en français et néerlandais ! Essayer la requête *~automobile -automobile* pour voir les mots trouvés en dehors du terme strict automobile. ~arabic retourne *Egypt*, *Lebanon*, *Arab* et … *Hindu* ! Plus de transparence dans la gestion des synonymes serait bienvenue.
- **"mot"** : Un mot exact. Google ne tient pas compte de l'accentuation pour la recherche mais favorise la forme spécifiée lors du tri des documents présentés.
- **"mot … mot"** : une suite de mots spécifiques, une expression
- **"mot * mot":** dans une suite de mots entre guillemets (et seulement là), une étoile peut être mise à la place d'un ou plusieurs mots complets qu'on ne désire pas spécifier. Par exemple : "ministère du * et du commerce"
- **site:www…** : un domaine d'origine. On peut être plus ou moins général et même indiquer des domaines de premier niveau.
  Par exemple : site:org OR site:com
- **title:"mot … mot"** : une suite de mots spécifiquement dans le titre du document (balise <title>…</title> et/ou première balise <h1>…</h1>)
- **+mot** : chercher ce mot même si c'est un mot vide dans la langue de l'utilisateur (`+de` en français par exemple) et le chercher en tenant compte des accents (`+dés` par exemple). Un « + » est assumé si un seul mot est cherché : `thé` seul est cherché comme si on avait tapé `+thé`. (Cette forme a donc une signification très différente de celle de Altavista où le "+" indiquait des mots obligatoires)

Lors du tri des documents, Google donne la préférence à la forme tapée : l'opérateur « + » n'a donc plus beaucoup d'intérêt.

- **mot-mot** : chercher un terme composé de plusieurs mots, qu'il soit écrit avec des tirets, des espaces ou même sans espace du tout: `gratte-ciel` trouve *gratte ciel*, *gratte-ciel* et *gratteciel* . `gratte-ciel` ne signifie pas du tout la même chose que `gratte -ciel` (voir l'opérateur "-"). Attention: ceci ne fonctionne bien qu'avec un seul tiret (`va-nu-pied` ne fonctionne pas bien !).

Opérateurs logiques (booléens):

- ***espace*** : les documents doivent contenir ce qui est à droite ET ce qui est à gauche. Le tri de Google favorise les documents où les différents mots spécifiés sont proches l'un de l'autre (voir plus bas).
- **OR** ou **|** : les documents peuvent contenir ce qui est à droite OU ce qui est à gauche. Attention : OR doit être écrit en majuscules !
- ***espace-*** (signe moins) : exclure les documents contenant le mot qui suit (SAUF)
- **(…)** : sous-expression à évaluer avant de faire les opérations avoisinantes

Le GoogleGuide vous donne d'autres exemples : http://www.googleguide.com/advanced_operators_reference.html

Le site de HotBot Etats-Unis fournit un formulaire de recherche Google parfois plus pratique que celui de Google même.

Limites:

- Les requêtes sont limitées à **32 mots**.
- Seuls les 1000 premiers résultats pertinents pour une requête sont accessibles, et ce même si les correspondances sont plus nombreuses. Les résultats peuvent même parfois être moins de 1000 en raison de la suppression des pages provenant d'un même site. D'après Google, obtenir plus de 1000 résultats entraînerait une lourde charge supplémentaire pour une demande finalement assez rare. Normalement, le tri assure que les références les plus utiles sont en premier (qui peut le vérifier ? les concurrents aux prises avec les mêmes problèmes techniques ?)

Dates:

- Lors d'une **recherche par dates**, la date est celle de l'indexation dans la banque de données (i.e. la visite du « spider » Google) et non celle de la publication effective de la page (telle que fournie par le serveur *http://*)
- Dans le formulaire de recherche avancée, vous pouvez faire une recherche sur les derniers 3, 6 et 12 mois.
- L'opérateur **daterange:***date julienne-date julienne* (ou le formulaire du site de HotBot) permet de spécifier un autre intervalle de dates. Une date

julienne est le nombre de jours écoulés depuis le début de notre ère : le site http://www.numerical-recipes.com/julian.html peut vous aider à le calculer.

Tri du résultat:

La qualité de Google vient de sa capacité à montrer en premier les pages jugées les plus pertinentes en général et les plus pertinentes à une recherche en particulier.

Google trie les documents trouvés en fonction:

- de **mesures de qualité** du site en général et aussi de chacune des pages (cohérence des méta-informations avec le texte visible de la page par exemple). Ces mesures ne sont pas ou peu documentées.
- une mesure du poids de chacune des pages indexées: Il s'agirait de l'algorithme PageRanks dont voici un extrait cité de Google :

  We assume page A has pages T1...Tn which point to it (i.e., are citations). The parameter d is a damping factor which can be set between 0 and 1. We usually set d to 0.85. There are more details about d in the next section. Also C(A) is defined as the number of links going out of page A. The PageRank of a page A is given as follows:

  PR(A) = (1-d) + d (PR(T1)/C(T1) + ... + PR(Tn)/C(Tn))

  Note that the PageRanks form a probability distribution over web pages, so the sum of all web pages' PageRanks will be one.

  PageRank or PR(A) can be calculated using a simple iterative algorithm, and corresponds to the principal eigenvector of the normalized link matrix of the web.

  Voir aussi: http://www.iprcom.com/papers/pagerank/
- d'un calcul de la **pertinence de la page** vis-à-vis de la recherche effectuée. Ceci se fait en tenant compte:
  - de la présence dans la page des mots de la recherche (éventuellement élargis à leurs synonymes ou à leurs variantes singulier/pluriel)
  - de l'emplacement de ces mots dans la page (titre, méta-données, texte) ou dans les liens vers cette page: ce dernier point cause parfois des problèmes éthiques car une page se retrouve indexée par les mots que d'autres personnes que ses auteurs utilisent pour la désigner. (Essayez: "*miserable failure*", l'auteur de la page visée ne cherchait pas consciemment ce qualificatif !)
  - pour chaque mot, du nombre d'occurrences mot dans la page pondéré par l'inverse de la fréquence relative de ce mot dans la partie du Web indexée par Google:

    $$w_i = tf_i * log\left(\frac{D}{df_i}\right)$$

- tfi = fréquence du terme **i** dans la page
- dfi = nombre de pages dans le Web contenant le terme **i**
- D = nombre de documents dans le Web

Cette formule a été mise au point par Gérard Salton (1927-1995), Université Cornell, sur base de la Théorie de l'information de Shannon.

- de la distance dans la page entre les mots cherchés: plus ils sont proches l'un de l'autre, plus la page est jugée pertinente vis-à-vis de la recherche effectuée.

Voir: http://www.google.com/librariancenter/articles/0512_01.html

- du **pays** indiqué par l'URL d'accès à Google : **google.be** accorde une nette préférence aux sites belges, **google.fr** aux sites français, **google.com** aux sites américains et **google.co.uk** aux sites anglais, etc. Il est réellement important de choisir la « localisation » de ses recherches. La page suivante devrait plus souvent servir de page de démarrage d'une recherche:
  http://www.google.com/language_tools?hl=fr
- de la **langue de l'utilisateur** qui est aussi celle des mots cherchés : le seul formulaire permettant de la spécifier est sur http://www.google.com/language_tools?hl=fr .
  Le seul autre moyen de changer la langue de l'utilisateur est de modifier « à la main » l'URL de Google (http://www.google.be/search?hl=fr&q=...) en changeant le paramètre **&hl=*xx*** (*xx* étant le code en deux lettres de la langue désirée).
  Il est essentiel de faire ses recherches en changeant sa langue d'utilisateur en fonction de la langue des mots cherchés. Google trie alors les documents en favorisant cette langue (et utilisera peut-être un jour le bon dictionnaire de synonymes). Il utilise alors l'algorithme adéquat pour rendre équivalents le singulier et le pluriel, le féminin et le masculin (rappel: le néerlandais semble mal supporté pour l'instant).

Le moteur de recherche Exalead.com

Exalead est le concurrent européen de Google, Nous n'avons pas encore déployé encore autant d'efforts que pour Google pour le tester et nous osons espérer qu'Exalead documentera bien son fonctionnement effectif.

La page d'accueil de la « Recherche avancée » récapitule bien les possibilités : http://preview.exalead.fr/search?password=beta. Avec le résultat d'une recherche, on trouve à droite des propositions automatiques de différents critères (termes fréquents, formats des documents, langues, etc.) pour affiner ses résultats.

Les expressions de recherche comportent les possibilités habituelles des systèmes de recherche documentaire :

- Expressions booléennes complètes (parenthèses pour les sous expressions, opérateurs AND, OR, AND NOT mais aussi NEAR)
- Troncature mais à droite seulement
- Recherche d'une phrase exacte, recherche phonétique, recherche d'un mot avec une éventuelle erreur de frappe

Termes à chercher:

- *H2SO3* est cherché comme un seul mot et Exalead ne trouve alors pas les documents avec *H 2 SO 3* ou H_2SO_3 dans leur texte. Ceux-ci sont retrouvés en demandant "H 2 SO 3". En théorie H-2-SO-3 devrait donner un bon résultat : une bonne question à poser aux développeurs de Exalead !
- Les mots "vides" (dans la langue de l'utilisateur) ne sont pas pris en compte (voir **+mot** plus loin). Thé ne trouve RIEN (« the » est un mot vide en anglais) et "thé" répond « Nos serveurs sont indisponibles » !
- *mot* : Un mot exact sans tenir compte de l'accentuation
- *mot****** : tous les mots commençant avec les caractères indiqués
- **"***mot* **…** *mot***"** : une suite de mots telle quelle, une expression
- **site:www…** : un site d'origine
- **intitle:"***mot* **…** *mot***"** : une suite de mots spécifiquement dans le titre
- **language:***xx* : restreindre aux documents de la langue spécifiée (fr : français, en : anglais, nl : néerlandais, etc.)
- **date >=** *aaaa/mm/jj* : les documents modifiés depuis la date indiquée
- **date <=** *aaaa/mm/jj* : les documents modifiés à la date indiquée ou avant
- */regexp/* : Expressions régulières s'appliquant sur un mot (et non pas sur une suite de mots ou sur de la ponctuation)

Dans une /expression régulière/ :

- le point (**.**) peut être utilisé à la place de n'importe quel caractère,
- ***** suffixe ce qui peut se répéter,
- **?** suffixe ce qui est optionnel
- **|** sépare des alternatives,
- les parenthèses peuvent être utilisées pour grouper des sous-expressions.

Par exemple :

- /j(unio)?r/ trouve aussi bien « Jr » que « Junior »
- mp(1|2|3|g) trouve les mp1, mp2, mp3 ou mpg
- /envel*op*e/ NEXT /ad*res*e*/ permet de trouver « enveloppe adressée » même avec des répétitions de lettres de mauvais aloi

Opérateurs logiques (booléens):

Toujours taper le nom de ces opérateurs en majuscules !

- *espace* (ET)
- *mot* **NEAR** *mot* : les deux opérandes doivent être à moins de 16 mots l'une de l'autre, dans n'importe quel ordre
- *mot* **NEXT** *mot* : idem, mais le deuxième mot doit apparaître après le premier
- *mot* **OR** *mot* (OU)
- **AND NOT** ou *–mot* (SAUF) : exclure un mot
- **OPT** *mot*: mot optionnel mais augmentant la pertinence du document
- **(…)** : une sous-expression logique (booléenne) peut être spécifiée à gauche ou à droite des opérateurs ET (espace), OU (OR), SAUF (AND NOT)

Tris des résultats:

Normalement par pertinence décroissante avec un algorithme très semblable à celui de Google car les résultats de l'un sont proches de ceux de l'autre. Comme Google, le tri favorise les documents où les différents mots spécifiés sont proches les uns des autres.

Le tri par date est aussi possible:

- **sort:new**
- **sort:old**

Stratégie de recherche documentaire

Procédure de recherche suggérée par l'Université de Berkeley:

1. What UNIQUE WORDS, DISTINCTIVE NAMES, ABBREVIATIONS, or ACRONYMS are associated with your topic?
These may be the place to begin because their specificity will help zero in on relevant pages.

2. Can you think of societies, organizations, or groups that might have information on your subject via their pages?
Search these as a “phrase in quotes”, looking for a home page that might contain links to other pages, journals, discussion groups, or databases on your subject. You may require the “phrase in quotes” to be in the documents’ titles by preceding it by **title:[no space]**

3. What other words are likely to be in ANY Web documents on your topic?
You may want to require these by joining them with **AND** or preceding each by **+[no space]**

4. Do any of the words in 1, 2, or 3 belong in phrases or strings -- together in a certain order, like a cliché?
Search these as a “phrase in quotes”. (E.g., “affirmative action” or “communicable diseases”)

5. For any of the terms in #4, can you think of synonyms, variant spellings, or equivalent terms you would also accept in relevant documents?
You may want to allow these terms by joining them by **OR** and including each set of equivalent terms in **( )**.

6. Can you think of any extraneous or irrelevant documents these words might pick up?
You may want to exclude terms or phrases with **-[no space] before each term,** or **AND NOT**

7. What BROADER terms could your topic be covered by?
When browsing subject categories or searching sites of webliographies or databases on your topic, try broader categories.

Procédure complète -- Besoins et solutions

Vous cherchez d'abord des documents qui évoquent vos besoins, sans forcément tout de suite mentionner les axes de solution: vous laissez ainsi un espace pour la "surprise", l'axe de solution que vous n'aviez pas imaginé.

La procédure ci-dessous est donc reprise plusieurs fois: pour des expressions de besoin et ensuite pour des axes de solution.

L'anglais et le français étant souvent semblables (plusieurs dizaines de milliers de mots similaires entre les deux langues), on a souvent l'illusion d'avoir couvert les documents dans les deux langues en une seule démarche. Il est recommandé de traiter chaque langue séparément en n'oubliant pas de changer la langue de l'interface utilisateur (paramètre **&hl=xx**).

Il y a des différences entre l'anglais « UK » et l'anglais « US » (centre vs center, globalisation vs globalization, lorry vs truck). L'opérateur « ~mot » (synonymes) de Google permet d'en tenir compte : Google devrait peut-être considérer certaines de ces différences comme des « variantes » comme le singulier et le pluriel.

Procédure complète -- Pour chaque expression de recherche

En une phrase courte mais précise et non ambiguë, pour quelle raison faites-vous une recherche ?

…

Quels sont les différents éléments (concepts, lieux, organisations, personnes, codes, etc.) mentionnés dans cette phrase ?

…

…

Pour chacun de ces éléments, écrivez une ligne dans un bloc note (NotePad) avec une suite de termes séparés par l'opérateur OR (en majuscules, abrégé par "|"):

…

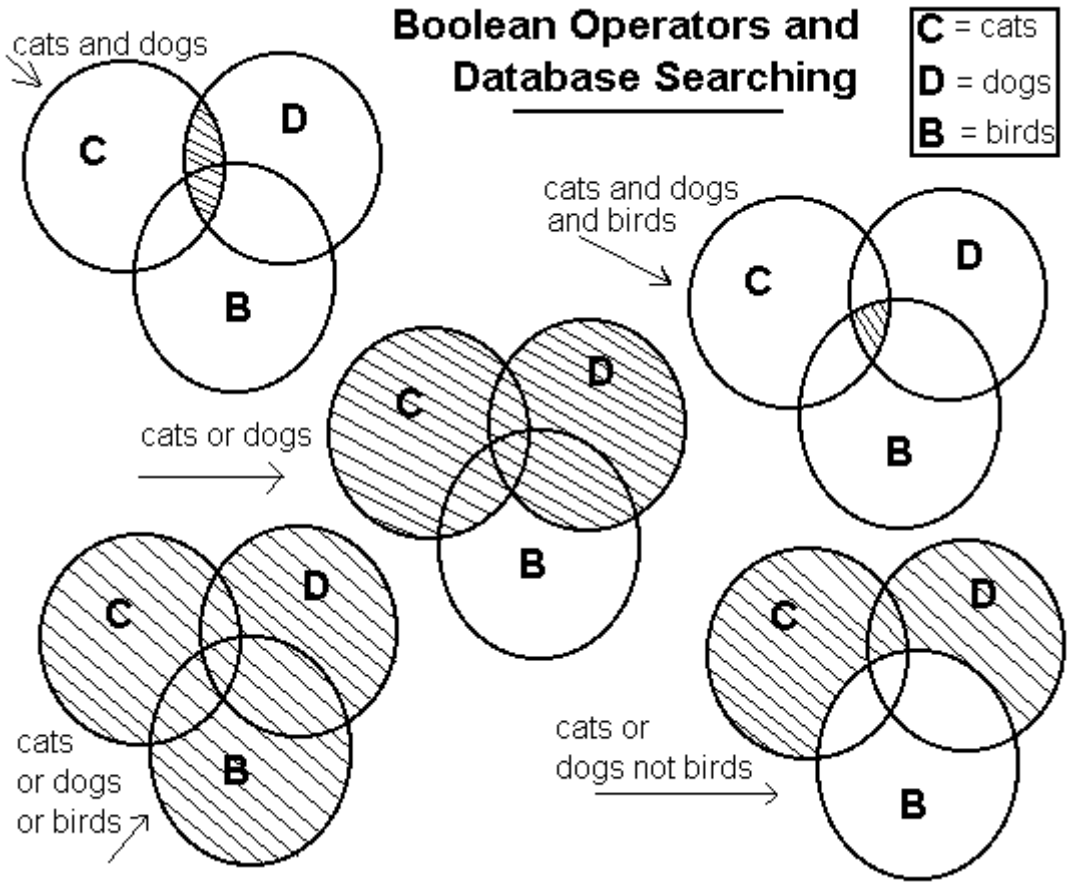
…

…

…

On donne plus loin différents conseils sur la préparation des différents types d'éléments

Petit rappel des opérations booléennes



Rappel et précision

Il s'agit pour une expression (besoins, axes de solution) de trouver un maximum de documents (rappel) avec un minimum de bruit (bonne précision).

- Pour chaque élément de l'expression de besoin ou de solution, on s'assure de l'exhaustivité en faisant **OR** (OU) entre chaque terme pouvant servir à nommer *précisément* cet élément (synonymes, spécifiques, antonymes complets, etc.). Exemple: baleine | baleineau | rorqual
  Attention : le tri de Google donne une préférence aux documents qui mentionnent plusieurs des différents termes liés par un **OR.**
- La précision sur un terme peut s'obtenir en contrôlant bien la distance permise entre les mots qui composent le terme et (le moins souvent possible) en rejetant les occurrences du terme si elles se trouvent près de mots indiquant un contexte incompatible. Exemple: baleine –parapluie
- Les parenthèses sont conseillées pour entourer chaque groupe de termes pour un concept (Google n'a pas l'air d'en faire grand-chose)
- La précision globale s'obtient en exigeant la présence de chacun des concepts présents dans l'expression de départ (espace (ET) entre les concepts)

Grandes étapes dans la préparation d'une recherche Google ou Exalead

Voici les principales étapes de la mise en œuvre de cette stratégie:

1. Identifier les **différents concepts** qui entrent dans l'expression du sujet de la recherche:

   Par exemple: *durée du congé de maternité*

2. Exhaustivité: Pour chacun de ces concepts, rassembler un **maximum de manière de l'exprimer** et faire un OU (union) entre chaque mot

   durée | longueur | semaines | mois

   Avec Google, on peut utiliser le tilde avant un mot pour qu'il mette lui-même des synonymes de ce mot:

   ~durée | longueur | semaines | mois

   Mais si l'expression qui exprime un concept est composée de plusieurs mots, il faut réunir ceux-ci par des guillemets ("congé de maternité") sans quoi on recevra aussi les pages où ces mots ne sont pas consécutifs.

3. Précision: Mettre **ces expressions de recherche d'un concept côte à côte** pour obtenir seulement les documents qui possèdent cette combinaison de concepts
   (ET implicite entre les groupes avec Google ou Exalead)

   (~durée | longueur | semaines | mois)
   (~congé | pause | ~vacances)
   (maternité | ~accouchement | naissance)

Adapter sa recherche au type d'information

Un objet "identifié":

Identifiant ou Code = "suite de lettres ou de chiffres à taper telle quelle"

Ce type d'information est le plus sensible aux problèmes lexicaux (frontières variables entre les mots, impossibilité de chercher des signes de ponctuation précis). Les moteurs de recherche obligent à spécifier manuellement toutes les variantes pour faire une recherche exhaustive : la maturité d'une technologie se mesure à l'attention portée aux détails qui affectent les utilisateurs…

- Un **document**: ISBN, ISSN: avec ou sans ponctuation, avec ou sans chiffre-preuve. Par exemple: `2748900375 | "2-7489-0037"`
- Un **produit**: Code utilisé par le fabricant, les distributeurs; formule chimique. Par exemple, le CAS (Chemical Abstract Service Number) est souvent très précis (on peut souvent omettre le mot CAS) : `"CAS 118-55-8"`

- Un **site Internet**: Celui-ci est identifié par un « URL ». l'URL peut être dans le texte (l'utilisateur pouvant le copier/coller dans la zone « adresse » de son navigateur) ou il peut être dans une balise HTML <A HREF=…>. Dans le premier cas, la recherche `"www domaine be"` fait l'affaire. Dans le second cas, Google permet de chercher `inurl:"www.domaine.be"` .
- Un **numéro de téléphone** et les différentes variantes dans le regroupement des chiffres. Par exemple, pour trouver tous les documents mettant en contact avec une grande firme à Bruxelles, on pourrait vouloir chercher son numéro de téléphone avec les différentes façons de l'écrire : IBM `"02 225 3333" | "02 225 33 33" | "2 225 3333" | "2 225 33 33" | "022253333" | "22253333"`
- Une date avec toutes ses variantes de forme, de langue, d'ordre ou de précision. Exemple: `"27 05 1958"|"1958 05 27"|"27 mai 1958"|"may 27th 1958"|"may 27 1958"|"27 may 1958"|"27 05 58"|"58 05 27"|270558|580527|27051958|19580527`

Problème des ponctuations parasites: dès qu'une balise HTML (même invisible) se glisse dans le mot ou dans le code, celui-ci est considéré comme scindé en deux mots.

Exalead permet les /expressions régulières/ (chère à Unix : voir plus haut) mais sur les mots seulement. Ceci ne permet donc pas de réduire les recherches sur un code dont les frontières de mots varient.

Google permet d'écrire `354-1096` pour exprimer `"354 1096" | "3541096"`. Malheureusement, on ne peut pas mettre plusieurs tirets de file: `354-10-96` ne fonctionne pas bien

---

Un "nom propre"

Le principe général est de rassembler toutes les variantes du nom lui-même (initiales par exemple) et des éventuelles parties, régions (lieux) ou filiales (entreprises) :

Un lieu:

Identifier : Synonymes et abréviations / Traductions / **Spécifiques**

Par exemple : `Belgique|Bruxelles|Anvers|Gand|Liège|Namur`

Les noms de lieux ont souvent des variantes selon la langue qui sert à les nommer (`Liège|Luik`, `Moscou|Moscow|Москов`). Plusieurs langues admettent les déclinaisons qui font varier la fin d'un mot: l'opérateur de troncature ***** est alors bien utile (Exalead seulement).

Exemple: Quels sont les exportations de la République de Macédoine ?

```
~importation (république | republic) (macédoine | macédonien | macedonia | macedonian | makedonia | "F Y R O M")
```

La translittération est l'opération permettant d'adapter l'écriture dans notre alphabet d'un mot provenant d'un autre alphabet. Elle tente surtout de préserver la prononciation. Cette technique permettrait aux moteurs de recherche de rendre accessibles plus de textes étrangers sur des lieux, des

personnes ou des entreprises dont on a entendu le nom. Par exemple, *Република Македонија* serait indexé à « Republika Makedonuja ».

Une institution ou une entreprise:

Identifier : Synonymes / Sigles / Traductions / Partie de/composé de / Changement du nom à travers l'histoire

Il n'est pas nécessaire de faire `"I B M" | IBM`: Google le fait automatiquement pour les mots de une lettre : il suffit de taper `"I B M"`.

Une personne:

Identifier : Abréviations, ordre/absence des prénoms, d'une initiale (anglo-saxons)

Chercher quelqu'un s'appelant "*Prénom Initiale Nom*":

```
"Prénom I Nom" | "P I Nom" | "Nom Prénom I" | "Nom P I"
```

Google cherche alors automatiquement aussi bien "P I" que "PI". Si PI est un mot fréquent (et introduit donc du bruit dans la recherche), vous pouvez écrire:

```
"Prénom I Nom" | "+P +I Nom" | "Nom Prénom I" | "Nom +P +I"
```

On peut aussi exiger la présence ou l'absence d'accents pour discriminer entre des textes de différentes langues: `+mélanie` par exemple ne trouve pas Melanie Griffith.

Un "nom commun"

Trouver toutes les façons d'exprimer un concept peut demander toute une recherche en soit:

- Effort de réflexion pour trouver des synonymes et leurs variations lexicales
- Traductions vers d'autres langues: http://www.systransoft.com
- Utilisation de dictionnaires ou de thésaurus en ligne. Mentionnons:
  - http://atilf.atilf.fr/dendien/scripts/tlfiv4/showps.exe?p=combi.htm;java=no; (Trésor de la Langue Française)
  - http://dico.isc.cnrs.fr/dico_html/ (français et anglais)
  - http://wordnet.princeton.edu/ (anglais)
  - http://thesaurus.reference.com/ (anglais)

- **`mot`** : suffit si les variantes singulier/pluriel+féminin/masculin+avec/sans accents, proposées par le moteur de recherche pour votre langue, pourront suffire (ne fonctionne pas bien en néerlandais)
- **`~mot`** : l'élément n'est pas réellement un élément central de la recherche mais permet de mieux cibler le genre de résultat désiré (par exemple: ~definition, ~comparaison, ~problem, etc.) et vous pensez que les synonymes généraux et pour l'anglais seulement que Google peut proposer feront l'affaire

- **"mot … mot"** : les mots spécifiés doivent être, dans l'ordre, les uns à cotés des autres. Au besoin, ne pas hésiter à écrire plusieurs variations de l'ordre ou dans les abréviations (OR entre chaque forme) La ponctuation ne compte pas !
- Avec Exalead, vous pouvez aussi utiliser:
    - `mot NEAR … NEAR mot` mots proches dans le texte, quel que soit l'ordre
    - `mot NEXT … NEXT mot` idem mais ordre significatif
- **(mot -motIdentifiantUnContexte)** : Une variante de signification peut être supprimée en excluant un contexte où elle apparaît. Par exemple, `baleine -parapluie`, `pied -"va-nu-pied"`
- **"mot"** : la recherche doit être faite sur une séquence, un code ou une variante précise
- **+mot** : tenir compte des accents (*dés, thé*). Ceci n'est plus très utile depuis que Google favorise toujours la forme accentuée que vous indiquez

Un objet

Identifier : Synonymes, abréviations / Traductions / Spécifiques+Génériques / Changement du nom à travers l'histoire

livre | imprimé | parchemin | publication | almanach | …

Un concept

Identifier : Synonymes, abréviations / Traductions / Spécifiques+Génériques / Changement du nom à travers l'histoire ou selon les auteurs (écoles de pensée)

Une source

Les sites Internet sont identifiés par un nom de domaine. Ce dernier va, de gauche à droite, du spécifique au générique.

Vous pouvez restreindre une recherche en indiquant *site:domaine*. Comme tous les niveaux de domaines sont permis, vous pouvez, par exemple:

- `… site:db.amazone.be` pour les banques de données de l'ASBL Amazone
- `… site:amazone.be` pour tout le site de l'ASBL Amazone
- `… site:qc.ca` pour un site québécois
- `… site:ca` pour un site canadien
- `… site:ac.be` pour un site académique belge
- `… site:co.uk` pour un site commercial du Royaume Uni

Vous pouvez faire | pour unir plusieurs domaines:

`… site:co.uk | site:com | site:biz` pour un site commercial

Un nombre

Google est capable de chercher sur un intervalle de nombre à la condition que ceux-ci soient notés dans la forme nord américaine (123 456.9999) et non pas européenne (123.456,9999).

On peut alors chercher sur un intervalle noté `minimum..maximum` (minimum et maximum pouvant avoir des décimales).

On peut aussi écrire `numrange:minimum-maximum` (minimum ou maximum peut alors être omis pour indiquer un intervalle ouvert).

On ne peut pas chercher des nombres négatifs (commençant par un "-") !

L'analyse du résultat des recherches

Les moteurs de recherche retournent un court extrait de chacun des documents trouvés, par pages de 10 à 20 documents.

Il est intéressant de repérer le type d'un document sans même cliquer pour le visualiser: les menus, index, tables des matières, répertoires, fichiers de données, etc. sont souvent inutiles mais retrouvés car ils contiennent un grand nombre de mots différents. Les moteurs de recherche ne sont pas encore capables de filtrer en fonction du type des documents.

Avec Firefox, il est pratique de parcourir la ou les pages de résultats en envoyant chaque page potentiellement intéressante dans un onglet différent: on n'est pas dérangé pendant le chargement des pages et celles-ci sont toutes prêtes quand on veut examiner leur contenu par la suite.

Le document:

Les créateurs des moteurs de recherche font déjà des efforts importants pour analyser la *forme* des documents pour en faire des critères de tri ou de recherche (balises donnant des meta-informations, textes dans des balises qui les mettent visuellement en valeur, textes dans des liens vers d'autres pages, comparaison des évolutions des documents entre deux passages, « profondeur » de la page dans le site, etc.).

Il serait intéressant que les moteurs de recherche permettent de spécifier des critères encore plus sophistiqués sur les documents. Voici quelques idées d'informations qu'ils devraient pouvoir récolter et structurer automatiquement.

Le contexte d'un document:

- Site: on a déjà l'opérateur `site:domaine` vu plus haut
- Auteurs (institutions, personnes)
- Hiérarchie et taille des unités documentaires. Est-ce un document complet ou un document découpé pour une navigation plus facile? La recherche pourrait elle être affinée en considérant comme des pages indépendantes les sections d'un grand document ?

- La ou les langues du document (une langue par unité documentaire?). On a déjà l'opérateur `language:xx`
- Les sujets normalisés selon un ou l'autre thésaurus (par exemple le MESH en médecine) ce qui ouvrirait la voie à la cartographie automatique des sujets d'un ensemble de document.
- Les formats de représentation:
    - HTML
    - Microsoft Power Point, Word, Excel
    - GIF, PNG, JPEG, MPEG
    - Adobe Acrobat PDF
    - etc.

  On a déjà l'opérateur `filetype:extension`

Les éléments d'un document:

- Méta-données
- Titre: on a déjà `intitle:"mots à chercher"`
- Menus de navigation
- Table des matières
- Texte
- Bibliographie

Les types de document:

- Menus de navigation
- Listes de liens
- Glossaires, dictionnaires, définitions
- Normes, standards
- Catalogues de cours
- Catalogues de produits
- Plaquettes publicitaires
- Documentations techniques de produits
- Cours
- Blogs
- Forum
- Images
- Enregistrements
- Biographies
- Critiques
- etc.

C'est ce type d'améliorations qui se dessinent avec les Digital Libraries, les Open Archives et le Semantic Web.

Bibliographie sur les méthodes de recherche

- **Aeris**, Aide aux étudiants pour la recherche d'information scientifique, Guillemette Lauters, 1999-2004, http://users.11vm-serv.net/aeris/
- **CERISE**, Conseils aux Etudiants pour une Recherche d'Information Spécialisée Efficace, URFIST de Paris, 1999, http://www.ext.upmc.fr/urfist/cerise/index.htm
- **Infosphère**, Apprendre à faire une recherche d'information efficace, Service des bibliothèques de l'UQAM, 2004, http://www.bibliotheques.uqam.ca/InfoSphere/
- **SAPRISTI**, des Sentiers d'Accès et des Pistes de Recherche d'Information Scientifiques et Techniques sur Internet, Doc'INSA, INSA de Lyon, 1997-2004, http://docinsa.insa-lyon.fr/sapristi/
- **University of California Berkeley Library** "Teaching Library Internet Workshops" http://www.lib.berkeley.edu/TeachingLib/Guides/Internet/About.html