Approfondir son expertise en recherche d'information

11-12 & 13 mars 2013

403-12

Conception et Animation : STEPHANE COTTIN – SGG (Services du Premier ministre)

SOMMAIRE

Objectifs

- Identifier les outils et méthodes de recherche à disposition de l'internaute aguerri
- Exploiter les fonctionnalités avancées de recherche
- Actualiser ses connaissances sur les moteurs de recherche et le web social

Méthode

Ce stage s'appuie sur l'alternance d'apports théoriques et d'exercices pratiques

Programme

1 Rappels et approfondissements....4
1.1 Spécificités de la recherche sur le web....4
1.2 Diversité des besoins, des contenus et des outils....6
1.3 Typologie des outils de recherche : moteurs généralistes et spécialisés, métamoteurs, annuaires généralistes et spécialisés, portails?....7
1.4 Typologie des sources : web visible, web invisible....11
2 Méthodologie....15
2.1 Comment s'organise la recherche d'information ?....15
2.2 Planifier ressources et moyens....17
2.3 Recourir à des experts / développer son expertise....18
2.4 Plan de recherche....19
2.5 Choix des mots-clés....23
2.6 Opérateurs avancés de recherche....26
2.7 Astuces pour identifier rapidement des sources d'information, des experts, un type de contenu spécifique, etc....29
3 Apport des outils et pratiques du web 2.0 : en quoi sont-ils créateurs de valeur ?....30
3.1 Archives ouvertes....30
3.2 Blogs, wikis?....32
3.3 Tags et folksonomie....34
3.4 Moteurs personnalisables....36
3.5 Partage de signets/présentations/images/vidéos?....37
4 Evaluer l'information sur Internet....38
4.1 Quelques questions clés à se poser....38
4.2 Comment évaluer un site web ?....38
4.3 Quelques outils pratiques....40
5 Autoformation....42
5.1 Veille sur l'actualité des outils de recherche d'information....42
http://sapristi-docinsa.insa-lyon.fr/guides-similaires....42
5.2 Veille sur les producteurs de sources d'informations....43

Contact et suivi : Stéphane Cottin : stephane.cottin@gmail.com
compte twitter : @cottinstef / autres coordonnées sur http://cottin.tel

http://www.adbs.fr/net-recherche-2010-le-guide-pratique-pour-mieux-trouver-l-information-utile-et-surveiller-le-web-82253.htm

Ouvrage épuisé. Parution de la nouvelle édition prévue courant 2013.

Net recherche 2010 : le guide pratique pour mieux trouver l'information utile et surveiller le web

Véronique Mesguich et Armelle Thomas. Préface d'Olivier Andrieu

Collection : Sciences et techniques de l'information

2010, 341 page(s), ISBN 978-2-84365-124-3

Chapitre 1 Diversité des besoins, diversité des contenus

1. L'internet, un univers complexe
2. La nouvelle physionomie du web
3. Une grande diversité de besoins
4. Moteurs de recherche web : des arbres qui cachent la forêt ?
5. Les dix règles d'or de la recherche d'information sur Internet

Chapitre 2 La recherche par mots-clés : les moteurs sacrés rois des outils

1. La lame de fond de l'approche mots-clés
2. Les moteurs de recherche : principes et idées reçues
3. Principaux moteurs français et internationaux
4. Les moteurs spécialisés, verticaux et personnalisables
5. L'exploration du web invisible : les moteurs gagnent du terrain
6. Les métamoteurs: innover ou mourir
7. Évolution des moteurs de recherche : dix tendances actuelles

Chapitre 3 Pour une recherche thématique : des annuaires généralistes aux portails spécialisés

1. La recherche thématique : l'information à la source
2. La grande famille des annuaires de recherche
3. Rechercher l'information économique et financière
4. Rechercher l'information scientifique et technique

Chapitre 4 L'apport du web social à la recherche d'informations

1. Une nouvelle approche du web pour un nouvel appétit d'échanges et d'action
2. Le partage de liens nouvelle génération
3. Outils de blogs et recherches dans la « tagosphère »
4. Réseaux sociaux : des millions d'amis
5. Une méthodologie de recherche « 2.0 » ?
6. Du web 2.0 au web 3

Chapitre 5 Net veille, la recherche automatisée

1. Les outils de surveillance du web
2. L'importance croissante des flux RSS
3. Les agents de surveillance
4. Les services de monitoring mots-clés

Chapitre 6 Commentaires de la méthode : des exemples de recherches

1. Les deux principales approches méthodologiques
2. Exemples de recherches détaillées pas à pas
3. Exemples de recherches « rapides »

Chapitre 7 Questions-réponses

1. Comment choisir ses mots-clés ?
2. Quels sont les opérateurs de recherche indispensables ?

...

En guise de conclusion

Google, toujours et partout, mais jusqu'à quand?
L'émergence de la recherche communautaire
Quel moteur idéal pour les utilisateurs ?
Le web sémantique pour renforcer l'« intelligence » des machines
Un web 3.0 déjà en route, un avenir très ouvert

Ouvrage épuisé. Parution de la nouvelle édition prévue courant 2013

Pourquoi un livre sur un thème aussi mouvant que la recherche d'informations sur Internet ? Parce que ce sujet en constante évolution reste finalement assez peu étudié dans son ensemble. Or il est indispensable de proposer aux internautes une vision globale qui leur donne le recul nécessaire face à la prolifération d'informations en ligne et leur permette d'identifier des contenus de qualité répondant précisément à leurs besoins. Sous l'apparente facilité d'utilisation des moteurs de recherche se cache en effet une réalité complexe, et le secret de la réussite d'une recherche ou d'une veille passe autant par la maîtrise des aspects techniques que par la capacité à évaluer et sélectionner les sources pertinentes.

1 Rappels et approfondissements

1.1 Spécificités de la recherche sur le web

1.1.1 Caractéristiques de l'information sur Internet

- ✓ Grande hétérogénéité dans les contenus et dans les publics (grand public et professionnels)
- ✓ Contenus dynamiques et renouvellement continuel
- ✓ Instabilité des localisations
- ✓ Fragmentation plus ou moins importante, selon les disciplines
- ✓ Multilinguisme et couverture géographique mondiale
- ✓ Information gratuite et payante (tendance à plus d'information, plus rapide, moins chère, avec une frange d'information à valeur ajoutée payante) : notion de freemium

1.1.2 La Taille du Web

Il est très difficile d'estimer la taille réelle du Web. Sa croissance se poursuit à un rythme excessivement rapide (on avait coutume de d'estimer à quelque 7 millions de pages supplémentaires par jour en 2002, certainement 10 fois plus en 2005, 100 à 1000 fois plus en 2012), mais de nombreuses pages ont une durée de vie très limitée. La plus grande difficulté provient aujourd'hui du nombre très important de pages dynamiques (cf le chapitre consacré au web invisible), et donc de la définition que l'on donne à une "page web".

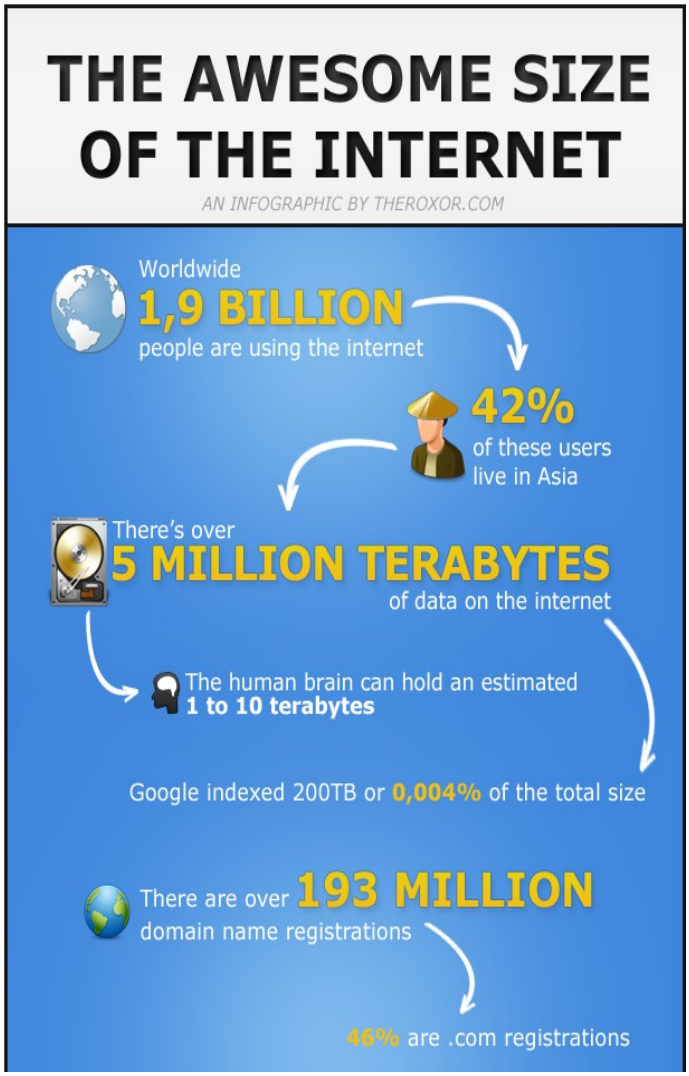
Cela dit, en toute logique, on doit dépasser actuellement les 1000 milliards de pages, sans compter les informations contenues dans les bases de données. Google a dépassé début 2005 les 8 milliards de pages indexées, date à laquelle il a cessé d'indiquer sur sa page d'accueil cette information, certainement erronée.

Cette question est très souvent posée et de nombreux débats en ressortent, qui permettent d'illustrer diverses évolutions récentes de la présentation des contenus sur internet .

Web bibliographie =

Application : une simple recherche sur google (FR) à l'expression « taille du web » donnera :

- http://googleblog.blogspot.fr/2008/07/we-knew-web-was-big.html Un billet du blog de Google datant de 2008 annonce 1000 milliards de pages (un « trillion US »)
- la notice wikipedia http://fr.wikipedia.org/wiki/World_Wide_Web sur le web
- la fiche yahoo answer se posant la question http://fr.answers.yahoo.com/question/index?qid=20101007024709AAHOa2U
- un article du blog d'Olivier Ertzscheid de 2007 http://affordance.typepad.com/mon_weblog/2007/03/question_de_tai.html
- un post du blog Brainsfeed (arrêté depuis) reprenant des informations de 2010 d'un site US proposant une infographie : http://urlpulse.co/blog/2010/10/28/the-awesome-size-of-the-internet-infographic/



Analyse : Même en posant une recherche sur un moteur censé répondre en français : on a très vite des réponses en anglais, via des articles de blogs reprenant les infos ailleurs. Phénomène de la « documentarisation » = tout est information, document, réutilisation possible (techniquement et juridiquement)

La présence de blogs et des encyclopédies collaboratives est évidente. Il faut apprendre à les maîtriser, ainsi que les autres méthodes du web 2.0 et du « crowd-sourcing »

1.1.3 La topologie du Web

Selon une étude menée par des chercheurs d'IBM, Compaq et AltaVista, parue en mai 2000, le Web aurait la forme d'un « nœud papillon » comprenant 4 parties. Le nœud ou "cœur" du net, très interconnecté, représentait 30 % des pages. Il est facile d'y accéder depuis de nombreux sites, simplement en suivant les liens. Environ 24 % des pages sont considérées comme "initiatrices". Leurs liens permettent d'accéder au cœur du web, mais la réciproque est fausse. À l'inverse, les pages "destination" (24 % des pages sondées) peuvent être facilement repérées depuis le cœur du web, mais elles n'y renvoient pas. Les 22 % restants sont des pages complètement disjointes du cœur. Elles peuvent être reliées à des pages initiatrices ou destination, voire même constituer des îlots totalement déconnectés. Il peut s'agir des pages

perso d'une famille ou d'un groupe d'étudiants, par exemple. Seule solution pour s'y connecter : connaître l'adresse, puisque même les moteurs de recherche ne peuvent les trouver.

Nombre de Dunbar
Théorie des 6 degrés de séparation, http://fr.wikipedia.org/wiki/Six_degr%C3%A9s_de_s%C3%A9paration
Règle des trois clics
Le « petit monde » de Milgram ou paradoxe de Milgram http://fr.wikipedia.org/wiki/%C3%89tude_du_petit_monde

1.2 Diversité des besoins, des contenus et des outils

1.2.1 Les besoins : Le cahier des charges du protocole IP et du Web ?

Rappel = le protocole TCP/IP (support du Net, créé fin des années 60), n'est qu'un protocole de communication parmi de nombreux autres. Il n'est pas considéré comme le meilleur, il a longtemps (et est toujours) concurrencé par de nombreux autres protocoles.
Son utilité a été avérée dans son utilisation comme support de messagerie par envoi de paquets d'information (maillage en réseau).
Le TCP/IP est d'abord un outil d'envoi de messages (e-mail ou SMTP), puis très vite, les listes de discussion.
L'usage des forums est apparu aussi très vite, contre l'avis des concepteurs, et c'est un autre protocole, l'USENET (ou NNTP) qui dès 1974 ouvrira les Usenet Newsgroups (ne pas confondre avec les google groups)
Voir http://boardreader.com pour avoir une idée de l'activité des forum (ou bulletin board system (BBS)), puis http://groups.google.com pour la recherche très ancienne ou par auteur http://groups.google.com/group/fr.doc.divers/browse_thread/thread/5cf919bd26beb50f/d0489619af5572f8#d0489619af5572f8
Enfin, et seulement début 1990, arrivent les outils d'affichage graphique propres au protocole TCP/IP. Tim Berneers Lee, et son équipe du CERN à Genève, crée le protocole HTTP (transfert hypertexte) et son langage de description de page (hypertexte markup language ou HTML), qui n'est qu'une version simplifiée du SGML des imprimeurs, uniquement conçue pour l'affichage sur écran : c'est donc une méthode pour les contenants. Mais l'html suffit à structurer superficiellement les données (les contenus), ce qui permet aux premiers robots de venir indexer (crawler) le réseau des réseaux (ou Web)

1.2.2 Les contenus : La construction et l'empilement des documents sur le Net

Dans l'ordre d'apparition :

- des messages ? Difficilement exploitables, mais énormément de données. Quand elles sont structurées, cela peut donner de très belles choses (projet Gutenberg)
- des forums ? Idem. L'absence de structuration et de métadonnées en font des ressources peu exploitables, sauf quelques domaines très pointus : l'informatique, et quelques sciences dures, mais aussi la recherche et l'analyse de rumeurs.
- des pages web ? La face émergée de l'iceberg, et pourtant déjà gigantesque
- du web structuré (sémantique) ! Arrivée peu à peu de « big data sets » = encore faut il savoir les réutiliser

1.2.3 Les outils : De l'adéquation aux ressources et aux besoins ?

http://outils.abondance.com/
Il faut des outils pour aller sur les sites : les navigateurs (Internet Explorer, Mozilla Firefox (ex Netscape, ex Mosaic), Google Chrome, Safari, Opera, etc.)

Des outils pour concevoir les sites : éditeurs html (Frontpage, Dreamweaver, Konqueror, Nvu, Blue Griffon...), Système de gestion collaborative de contenu web (CMS type EzPublish, Spip, Joomla, Typo3, Alfresco...)

Des outils pour analyser ces sites et ces contenus : moteurs internes ou externes, lecteurs de fils rss, aspirateurs, agrégateurs, …

Des outils de recherche : les moteurs et les annuaires :

1.3 Typologie des outils de recherche : moteurs généralistes et spécialisés, métamoteurs, annuaires généralistes et spécialisés, portails?

1.3.1 Les moteurs

Palmarès des moteurs de recherche (décembre 2011) en France
http://www.journaldunet.com/ebusiness/le-net/top-5-moteurs-de-recherche/
http://www.les-infostrateges.com/actu/11121337/le-top-5-des-moteurs-de-recherche-en-france

1. Google : "couverture active" (part d'audience) = 85 %
2. Google images 36,7 %
3. Voila 15,9 %
4. Bing 15,8 %
5. Ask 14,7 %
6. Yahoo

Extrait du sommaire de Net-recherche sur les moteurs
3.1 Quatre générations de moteurs
3.2 Disparitions et regroupements : un rapide historique
3.3 Moteurs de recherche : de l'oligopole au duopole
3.4 Google toujours loin devant
3.5 Yahoo! et Bing, les challengers
Yahoo!
Bing
3.6 Ask, Voila/Le Moteur/Orange, Exalead, et les autres…
Ask, Cuil . Exalead, Gigablast, Voila/Le Moteur/Orange

3.7 Une tendance à la normalisation des fonctionnalités
Le cache
Les suggestions de mot-clé
La page d'accueil personnalisable
Importance de Wikipédia dans les premiers résultats
Syntaxes d'interrogation : constantes et spécificités
Comparaison des fonctionnalités avancées de recherche
(…)

7. Évolution des moteurs de recherche : dix tendances actuelles 86
7.1 Simplifier la syntaxe et aider l'utilisateur 86
7.2 Permettre une recherche « universelle » 89
7.3 Personnaliser son moteur de recherche 90
7.4 Exploiter les technologies de clustering et de cartographie 92
7.5 Rechercher en langage naturel 96
7.6 Rechercher dans les fichiers audio et vidéo 98
7.7 Rechercher depuis les « smartphones » 101
7.8 Organiser les résultats sous forme de tableau 102
7.9 Traduire de manière toujours plus efficace 102
7.10 Permettre une recherche « temps réel » 102

Méthode de fonctionnement et critères de comparaison des moteurs de recherche

* Provenance de l'index, taille de l'index, ressources prises en compte
* Délai moyen de rafraîchissement et conditions de mise à jour
* Mode d'indexation et traitement éventuel des ressources (linguistique, statistique, parsing : extraction des éléments signifiants)
* Options de recherche simple et avancée, aide à la reformulation des questions.
* Critères déterminants pour le classement des résultats (tri de pertinence)
* Présentation des résultats : informations disponibles, source du résumé, datation des résultats, regroupement des pages d'un même site (cluster), mise en exergue des mots-clés sur la page, archive de la page, cartographie, etc.
* Critères subjectifs : interface de consultation, adéquation aux types de recherche effectués.

Le tri de pertinence des moteurs

Principes

Les moteurs mettent au point des "tris de pertinence" pour classer de façon automatique leurs résultats de recherche, afin de présenter en début de liste ceux qui obtiennent le meilleur score pour une requête donnée. Les algorithmes de tri sont différents en fonction des outils et plus ou moins performants et complexes. Ils ne sont généralement pas connus de façon précise et varient dans le temps pour chaque moteur. Les principaux critères utilisés sont les suivants :

* ***Par rapport à la requête de l'internaute*** **:**
  - position des mots dans la requête : Ainsi, sur Alta Vista et Google, l'ordre des mots de la question n'est pas neutre.

- correspondance d'expression : similarité entre l'expression de la requête et l'expression correspondante dans un document

* ***Par rapport aux pages de résultats***

- "densité" des mots-clés : nombre d'occurrences du (des) terme(s) demandé(s) / nombre de termes de la page en question, une fois éliminés les mots vides.
- présence dans le titre ou dans le premier tiers de la page
- mise en exergue du texte (gras, taille des caractères)
- présence dans les méta-données* (ce critère tend à perdre de son importance). Des outils comme Google ou Fast n'utilisent pas du tout ce critère, et Voila ne leur donne plus beaucoup d'importance.
- présence dans l'adresse de la page
- proximité des mots-clés sur la page

* ***Par rapport à la base de données du moteur :***

- rareté des mots (déterminé par le nombre d'occurrences du mot dans l'index) : des mots rares dans une requête ont une pondération plus importante que des mots communs
- popularité des pages : indice de clic (basé sur l'audience) ou indice de popularité (basé sur le principe de citation).

La popularité comme mesure de pertinence

Depuis quelques années, on a assisté à la naissance, au développement, puis au franc succès de deux nouvelles mesures de pertinence appelées respectivement "indice de clic" et "indice de popularité". Ces mesures s'ajoutent le plus souvent à d'autres "ingrédients" pour classer les résultats des moteurs, mais ils constituent aussi le critère de tri primordial des nouveaux venus inventeurs de ces technologies. Ces nouveautés, issues du "filtrage collaboratif", sont symptomatiques d'un certain désarroi des acteurs et utilisateurs du réseau face aux multiples difficultés d'un recueil rapide d'informations pertinentes.

* L'indice de clic

Il s'agit ici d'analyser le comportement des internautes posant la même question au moteur et de privilégier dans le classement les pages les plus "cliquées", et sur lesquelles le temps passé est le plus important. Il permet donc de classer les résultats des requêtes les plus populaires, en récupérant le jugement implicite de communautés d'usagers. Fonctionne donc en "tâche de fond" sur un moteur existant, la base s'enrichissant ainsi.

Direct Hit, racheté par Ask Jeeves en 2001, puis devenu Teoma, était la référence dans ce domaine et fut utilisés par de nombreux moteurs comme Lycos et MSN (plus de 50 sites clients), mais aussi Ask Jeeves. Alta Vista et Inktomi ont développé leur propre système sur un principe similaire. Mais Ask Jeeves a ensuite décidé de centraliser ses efforts de développement sur le moteur Teoma.

* L'indice de popularité

On s'intéresse ici aux "backlinks" ou "liens à l'arrivée", c'est à dire au nombre et à la qualité des liens pointant sur une page : on mesure ainsi sa popularité, et donc selon les concepteurs de ces technologies, sa pertinence. Les anglophones disent pour mieux expliquer le principe de l'indice de popularité :"It's not what you know, it's who knows you". En d'autres termes, le plus important n'est pas ce que vous dites ou ce que vous savez, mais qui vous connaît.

Le principe, rendu célèbre par le moteur Google, n'est pas totalement nouveau. Ne mesure-t-on pas la crédibilité d'un auteur scientifique au nombre de citations qui sont faites sur ses articles ?

Google examine la structure des liens sur l'ensemble du web. Quand on fait une recherche, un URL avec un fort "page rank" a plus de chance d'être listeée en premier. Chaque page de l'index de Google est notée : le "page rank" est une propriété de la page en elle-même, indépendante des requêtes effectuées : elle équivaut à la probabilité quel 'internaute aboutisse à cette page sur Internet.

Le tri des résultats pour une requête intègre d'autres critères plus classiques, dont bien entendu la présence des termes de la requête dans les pages de résultat, ou identifiée comme pertinente via l'analyse du contexte des liens.

Le grand avantage du système est de donner une meilleure visibilité aux sites incontournables du domaine de recherche. L'inconvénient majeur est là encore, de pénaliser les nouveaux venus peu connus.

Les sociétés spécialisées dans le référencement cherchent bien entendu à connaître le plus précisément possibles les critères clés de chaque moteur. L'objectif est de faire apparaître en bonne position (ranking) les pages web de leurs clients sur les listes de résultats à une requête comportant certains mots-clés.

Ce travail de référencement se fait parfois au mépris de l'éthique et donne lieu à une activité de "spamdexing" ou "spamming". (Création ou modification d'un document avec l'intention de tromper un catalogue ou un système de classement électronique. Toute technique qui a pour objectif d'augmenter la position potentielle d'un site aux dépens de la qualité de la base de données du moteur de recherche. Définition issue du glossaire réalisé par les membres francophones de la liste de diffusion I-Search Digest hébergé par le fournisseur d'hébergement IDF www.idf.net/mdr/glossaire.html).

Voir les sites spécialisés dans le référencement comme Webrankinfo (www.webrankinfo.com) qui donne de précieux conseils, notamment pour Google.

A noter : Le modèle du positionnement payant s'est aujourd'hui imposé, et tous les moteurs proposent des "résultats sponsorisés" (c'est aujourd'hui souvent l'indice de clic qui est favorisé) à ne pas confondre avec les résultats "normaux"

1.3.2 Les répertoires de recherche

Principe des répertoires de recherche

* "Collections" généralistes ou spécialisées de sites web classées par catégories organisées hiérarchiquement (le H de Yahoo signifie « hierarchically » Yet Another Hierarchically Organized Oracle)

* Filtrage et classement “ manuels ” : la sélection peut être plus ou moins rigoureuse, avec une évaluation et une description des sites éventuellement enrichies.

* Pas d’indexation en texte intégral des pages des sites.

* Outils de première approche : Donnent une vue d'ensemble d'un domaine à l'utilisateur, qui peut ensuite naviguer à l'intérieur des sites indiqués pour aller plus loin.

* Ne gèrent pas les requêtes complexes, mais permettent généralement de faire une recherche par mot-clé sur une catégorie seule.

* Problèmes de mise à jour et de “ désherbage ”.

Modes de recherche

* Recherche dans le plan de classement : Cette méthode est parfois complexe, aucune norme n'existant pour l'arborescence des répertoires. Les sites sont indiqués par ordre alphabétique.

* Recherche par mot clé : la recherche se fait sur les champs suivants : intitulés des catégories, titres des sites, résumé des sites, adresses URL des sites. Avec ce mode de recherche, les résultats bénéficient généralement d'un classement de pertinence opéré uniquement sur les fiches descriptives des sites. L'Open Directory ne recherche pas sur les catégories.

Utilisation

Les répertoires sont à réserver pour des recherches plutôt thématiques, ou sur des mots clés assez généralistes ; notons toutefois que les catégories deviennent au fil du temps de plus en plus "pointues" en fonction du sujet.

Si l'on utilise des mots clés trop précis, ou trop de mots clés, la plupart des répertoires passent le relais à des moteurs de recherche partenaires qui effectuent des recherches sur le texte intégral des pages web.

Les répertoires sont aussi utiles :

* pour se faire une idée du vocabulaire utilisé dans un domaine
* pour retrouver, à partir d'un site web donné, d'autres sites traitant du même sujet
* pour trouver des sites fédérateurs ou portails spécialisés
* pour obtenir rapidement tous les sites d'une organisation importante.

1.4 Typologie des sources : web visible, web invisible

1.4.1 Le web invisible

Il s'agit de l'ensemble des pages non localisables et/ou non indexables par les outils. Le web invisible correspond à plusieurs types de ressources :

- ✓ Pages dont les caractéristiques techniques rendent difficiles, sinon impossible l'indexation par les moteurs : frames, javascripts modifiant le contenu, technologies propriétaires.
- ✓ Pages qui n'ont fait l'objet ni d'un référencement direct, ni d'aucun lien d'une autre page.
- ✓ Pages nécessitant une identification de la part de l'internaute
- ✓ Pages dont le contenu indique aux moteurs qu'ils ne doivent pas l'indexer
- ✓ Page produite à partir de bases de données ou d'applications, et dont l'URL comporte des paramètres non exploitables par la plupart des moteurs
- ✓ Page produite à partir de données saisies par l'utilisateur via un formulaire html. Exemple : les résultats de l'interrogation d'une base de données avec des critères de recherche entrés par l'utilisateur.

(définition mise au point par les formateurs internet ADBS)

On ne connaît pas du tout la taille du web invisible

Une certitude : le web invisible croît plus rapidement que le web visible, du fait de la multiplication des bases de données à interface web, et de l'explosion du web dynamique.

A noter : les fichiers pdf ou flash, autrefois partie intégrante du web invisible, sont aujourd'hui indexés par plusieurs moteurs, Google en tête.

1.4.2 Informations de base sur les méta-données

Il s'agit au départ de balises du langage html qui permettent de donner des informations (description, mots-clés) sur le contenu d'une page web.

Elles se trouvent dans l'en-tête HTML de la page Web, (le "HEAD") et fournissent des informations qui ne sont pas visibles par les navigateurs. Les méta-tags les plus courants (et les plus utiles pour les moteurs de recherche) sont KEYWORDS (mots-clés) et DESCRIPTION.

Pour visualiser les méta-tags : Affichage Source (Explorer) / CTRL U (Firefox)

Le méta-tag KEYWORD permet à l'auteur de souligner l'importance de certains mots et phrases utilisés ou non dans sa page. Certains moteurs de recherche tiendront compte de cette information - d'autres l'ignoreront. Certains moteurs donneront en plus un « coup de pouce » dans le classement pour certains documents au cas où le mot clé de requête se trouve dans les méta-tags, mais ils peuvent pénaliser une page où un terme est répété plusieurs fois dans la balise meta keyword..

Le méta-tag DESCRIPTION permet à l'auteur de contrôler le texte affiché quand la page paraît au niveau des résultats d'une recherche. Certains moteurs de recherche peuvent ignorer cette information. Contrairement à KEYWORDS , DESCRIPTION est en langage naturel.

Pour pallier la "faiblesse" des balises méta classiques, certains groupements travaillent à mieux décrire les documents sur Internet. On pourra utilement se référer au "Dublin Core", métadonnée de 15 éléments destinée à la description générale des documents, qui est d'ores et déjà utilisée via les balises méta par certains organismes, y compris en intranet et normalisé dans la norme ISO 15836 http://fr.wikipedia.org/wiki/ISO_15836 . Le Dublin Core, considéré comme un bon candidat pour une norme internationale, est le fruit du travail depuis 1995 d'une cinquantaine de chercheurs et professionnels issus du monde de la documentation et des bibliothèques, de l'informatique, de la codification des informations. L'ensemble fut initié par l'OCLC (Online Computer Library Center) en accord avec le NCSA (National Center for supercomputing applications). Le Dublin Core doit son nom à la première réunion de travail en juin 95 à Dublin Ohio dans les locaux de l'OCLC.

http://fr.wikipedia.org/wiki/Dublin_Core

Élément	Élément (anglais)	Commentaire
1. Titre (métadonnée)	Title	Titre principal du document

2. Créateur (métadonnée)	Creator	Nom de la personne, de l'organisation ou du service à l'origine de la rédaction du document
3. Sujet (métadonnée) ou mots clés	Subject	Mots-clefs, phrases de résumé, ou codes de classement
4. Description (métadonnée)	Description	Résumé, table des matières, ou texte libre. Raffinements : table des matières, résumé
5. Éditeur	Publisher	Nom de la personne, de l'organisation ou du service à l'origine de la publication du document
6. Contributeur	Contributor	Nom d'une personne, d'une organisation ou d'un service qui contribue ou a contribué à l'élaboration du document. Chaque contributeur fait l'objet d'un élément Contributor séparé
7. Date (métadonnée)	Date	Date d'un évènement dans le cycle de vie du document
8. Type de ressource	Type	Genre du contenu
9. Format	Format	Type MIME, ou format physique du document
10. Identifiant de la ressource	Identifier	Identificateur non ambigu : il est recommandé d'utiliser un système de référencement précis, afin que l'identifiant soit unique au sein du site, par exemple les URI ou les numéros ISBN. Raffinement : Is Available At
11. Source	Source	Ressource dont dérive le document : le document peut découler en totalité ou en partie de la ressource en question. Il est recommandé d'utiliser une dénomination formelle des ressources, par exemple leur URI
12. Langue (métadonnée)	Language	
13. Relation (métadonnée)	Relation	Lien avec d'autres ressources. De nombreux raffinements permettent d'établir des liens précis, par exemple de version, de chapitres, de standard, etc.
14. Couverture (métadonnée)	Coverage	Couverture spatiale (point géographique, pays, régions, noms de lieux) ou temporelle
15. Droits (métadonnée)	Rights	Droits de propriété intellectuelle, Copyright, droits de propriété divers

Un 16e élément apparaît parfois, l'Audience, mais il ne figure pas dans la liste de la norme ISO 15836.

2 Méthodologie

2.1 Comment s'organise la recherche d'information ?

2.1.1 Les dix règles d'or de la recherche d'information sur Internet

1. **"Affiner"** savoir poser les bonnes questions :sa question (type de recherche, sujet précis et objectif, étude des concepts, recherches préliminaires éventuelles), choisir ses stratégies de recherche. (OA "lorsqu'on a une recherche à faire sur le web, la première chose à faire, c'est de ne pas aller sur le web")
2. **Maîtriser** les outils de navigation et de recherche : gestion des signets, récupération des données, répertoires, moteurs et méta-moteurs. Pour les moteurs, utiliser au moins deux moteurs ayant des approches différentes et complémentaires.
3. **Trouver** de bons points de repère : annuaires et "bons sites" (associations professionnelles, experts, usuels du domaine) dans un domaine :
    - Retrouver les équivalents de ses sources habituelles (d'où l'importance d'avoir une idée, même approximative, de l'offre documentaire dans le domaine recherché).
    - Compléter avec les sources originales
    - Trouver les répertoires et "méta-pages" spécialisées.

    Une adresse fiable qui renvoie directement au sujet d'une recherche constitue un bon point de départ parce que :
    L'administrateur d'un bon site spécialisé est généralement averti de l'existence et la création des autres sites de la spécialité : Il sélectionne les meilleures références et parfois les commente ; Il passe du temps sur le réseau dans son domaine de compétence ; Il met en jeu son expertise.
4. **Toujours analyser** l'information : recouper l'information, faire preuve d'esprit critique, évaluer rapidement
5. **Utiliser** en cours de recherche son carnet d'adresses pour **garder trace** des sites ou pages intéressants mais momentanément hors sujet, et "noter" rapidement les ressources enregistrées.
6. **Savoir se limiter** dans le temps : ne pas se rendre esclave d'une recherche d'exhaustivité à tout prix, ne pas s'obstiner en vain. Internet contribue souvent à répondre à la question "où trouver" (chercher l'info qui conduira à l'info).
7. **Choisir** les bons mots-clés
8. **Rester clair** sur ses objectifs, sa stratégie et ses critères de choix établis auparavant face à "l'hyper-choix". Rester vigilant sur la trajectoire parcourue et celle qui reste à parcourir. "on ne doit pas rechercher l'info de la même manière suivant que l'on est novice ou expert sur un sujet.
    Le novice recherche les sites web les plus riches et les plus visités. Il n'a pas de temps à perdre et veut éviter le bruit. Il obtient des résultats rapides, après la phase d'acclimatation au problème.

L'expert n'est pas intéressé par les sites classiques. Il recherche au contraire le bruit afin de trouver le "signal faible" qui lui donnera l'avantage. Il est prêt à y consacrer beaucoup de temps. (il fait beaucoup d'efforts pour des résultats marginaux)

9. **Conjuguer harmonieusement** recherche dans les outils classiques, web invisible, presse et actualité et navigation hypertexte : la recherche d'information sur Internet est un processus itératif qui oblige à passer par différents modes d'accès à l'information.
10. **Etre "agile"** : développer une lecture rapide, lancer plusieurs recherches à la fois, savoir rebondir d'une information à l'autre, d'un outil à l'autre, d'un article à une institution. Se souvenir qu'il n'existe pas de méthode infaillible et que chercher l'information sur Internet, c'est avant tout un état d'esprit. Ainsi, si je cherche le premier producteur de statistiques en Irlande, je peux commencer, sans trop de risques d'erreurs, par faire l'hypothèse que l'INSEE propose des liens vers ses homologues européens.

2.1.2 Problématique : *Faut-il commencer une recherche sur Internet ?*

Internet est-il complémentaire à d'autres supports ou se suffit-il à lui-même ? . On trouvera rarement matière à une étude complète d'un sujet via Internet (test : essayez avec un sujet que vous connaissez bien = vous serez toujours très déçu). Par contre, bien (et rationnellement utilisé) le Web sera souvent plus rapide et moins cher que d'autres supports pour des recherches de type "questions-réponses".
Enfin, Internet et ses différents services (mail, newsgroups, mailing lists) se prêtent bien à la pratique de la veille, de part leur caractère mouvant, décloisonné, international.

2.2 Planifier ressources et moyens

2.2.1 Critères :

- Temps
- Coût / abonnement
- Confidentialité (ressources externes / internes)

2.2.2 Quand utiliser quels outils ?

La réponse à cette question ne peut pas être définitive. Rappelons que la recherche d'information sur Internet n'est pas une science, et tout dépend aussi de son expérience de la recherche et du Web, et de sa façon de travailler.

Disons en simplifiant beaucoup…

En fonction du type de recherches

* Recherches larges ou première approche : Öannuaires généralistes
* Recherche d 'information ponctuelle (tous secteurs) : Ömoteurs généralistes
* Recherche sur des données de nature bien définie (statistiques, pays, presse, indicateurs…) : Öannuaires et outils spécialisés sur ce type de recherche
* Recherches récurrentes sur un sujet: Öidentification de sites via pages de liens ou annuaires spécialisés, puis recherche par navigation / Öméta-moteur off-line
* Recherches précises sur noms ou chaînes de caractères (sans booléens) : Öméta-moteurs.

En fonction de sa connaissance du sujet :

	Faible connaissance du sujet	Bonne connaissance du sujet
"Question-réponse"	.Recherche sur les moteurs ou méta-moteurs .Remonter à un concept plus généraliste et utiliser les annuaires	."Sites de référence" (Sites spécialisés sur le sujet, repérés au préalable)
"Tout savoir sur"	.Annuaires pour identifier les bons sites et les bons mots clés .Recherche sur " sites de référence" .Recherche sur moteurs	. "Sites de référence" complétés par recherches sur moteurs ou méta-moteurs .Personne référence.

2.3 Recourir à des experts / développer son expertise

Usage des listes de discussion

Exemple :

Liste ADBS-INFO Subject : INFO : Veille avec Google

From: christine griset <christinevince@live.fr>
Date: Thu, 22 Dec 2011

Assurer une veille professionnelle avec Google (JDN High Tech Comment ça marche du 22/12/11)

Veille concurrentielle, Google News, Google Alertes, Google Reader, iGoogle pour centraliser les informations

http://www.commentcamarche.net/faq/14175-assurez-une-veille-professionnelle-avec-google?utm_source=benchmail&utm_medium=ML32&utm_campaign=E10213573&f_u=18135858

cordialement,

Christine Griset, websurfer/

Accès aux archives de la liste : https://listes.adbs.fr/sympa/arc/adbs-info

Application : savoir chercher dans des archives de listes, (s'abonner à une liste, participer à une liste, etc.)

2.4 Plan de recherche

Principes d'une veille efficace sur Internet

Dire que l'on "fait de la veille sur Internet" est un abus de langage. En fait, on utilise Internet comme un outil de surveillance des entreprises, des marchés, des technologies, des évolutions de la société…

L'apport d'Internet par rapport dans une démarche de veille :

è Une information ouverte, disponible à tout moment, souvent à faible coût

è Une information régulièrement actualisée

è Un très grand volume d'information à disposition

è Des informations multi-sources, multidisciplinaires (le fonctionnement réseau étant idéal pour la veille).

è Une information numérisée, pouvant être triée et exploitée rapidement.

Mais il ne faut pas oublier les aspects négatifs :

èRisque de désinformation : une information "orientée" et donc pas toujours fiable.

è Risque de se "noyer" dans l'information.

èUne information parfois difficilement accessible (barrières des langues, services payants,...).

è Une information en perpétuelle évolution et donc instable

è Une relation temps-coût / valeur intrinsèque de l'information obtenue pas toujours facile à maîtriser.

Voir aussi chez François Magnan http://www.francoismagnan.info/ une de ses présentations disponibles sur Slideshare : http://fr.slideshare.net/FranoisMagnan/recherche-et-veille-documentaire

2.4.1 Méthodologie à mettre en œuvre

* **Définition des cibles de veille**

La mise en place d'un processus de veille sur Internet s'appuie sur un ciblage de la veille défini à partir des objectifs et du positionnement stratégique de l'entreprise ou organisation sur ses différents marchés.

Concrètement, c'est la réponse aux questions : Qui surveiller sur Internet ? Sur quel thème ?

* **Inventaire des sources connues sur Internet**

Lesquelles sont pertinentes par rapport à l'étape précédent, pour quel thème ?

* **Recherche d'autres sources pertinentes**

Pour cette étape, on procédera d'abord à la constitution évolutive d'une liste arborescente des mots-clés des différents thèmes stratégiques, traduits en anglais, et si nécessaire, dans d'autres langues.

Cette liste peut évoluer en fonction des ressources trouvées, et de l'évolution du vocabulaire du domaine.

Il s'agit ensuite de constituer les équations de recherche les plus pertinentes pour chaque thème de veille pour les proposer à différents moteurs.

On peut aussi travailler à partir de répertoires hyper-spécialisés et suivre les liens proposés (les répertoires généralistes sont de peu de secours, les thèmes de veille étant généralement assez pointus).

* **Mise sous surveillance des couples "ressource Internet"/ thème de veille**

On obtient donc une liste de ressources clés sur Internet qui pourra évoluer dans le temps (ne pas oublier les forums et listes de diffusion).

Après un choix d'agents à utiliser (agent d'alerte on-line ou off-line), les pages clés (par exemple pour un concurrent les pages Produits, News et Offres d'emploi) sont mises sous surveillance automatique.

Les équations de recherche peuvent être soumises régulièrement aux moteurs de recherche sélectionnés (voire méta-moteurs) pour être averti de la présence de nouveaux acteurs intéressants.

L'utilisation parallèle de logiciels de cartographie sur les résultats de ces requêtes, (téléchargés préalablement sur le disque dur) peut permettre de repérer des évolutions faibles ou tendances sur des marchés mouvants.

Avec ces outils, il peut être intéressant de travailler en plus sur des thèmes de veille élargis.

* **Collecte et Sélection des informations recueillies**

Rappelons que dans une optique de veille, on ne se base pas sur des données rétrospectives, ni même quantitatives et certaines, mais sur des signaux fragmentaires dits "faibles" : en ne conservant que les informations réellement stratégiques pour l'entreprise, la sélection consiste à affiner le travail de collecte et permet l'analyse.

L'évaluation de la fiabilité de la source et de l'information sont bien sûr très importantes, mais peuvent se faire a posteriori.

On quitte alors le "cycle Internet" pour intégration des données dans le système d'information de l'entreprise, diffusion et exploitation.

2.4.2 La veille automatisée

On a vue la richesse d'internet pour la mise en œuvre d'une veille. Cependant, l'exploitation manuelle est souvent délicate du fait de tâches très consommatrices en temps. Les outils (agents, cf "Les agents évolués sur Internet" ci-dessus) permettent une automatisation de tâches répétitives :

* Outils de collecte (moteurs d'indexation et de recherche, méta-moteurs, agents d'alerte) : ils permettent de surveiller des pages et des sites web (voire des dossiers de pages web), des catégories d'un répertoire, différentes catégories de ressources (actualités, articles de presse, appels d'offre, communiqués de presse, informations financières, etc.). Ils peuvent travailler sur un moteur, un répertoire, plusieurs outils simultanément, une base de données, voire plusieurs bases de données.
* Outils de tri et d'aide à l'analyse (résumés, traduction, text-mining, cartographies, etc.)
* Outils d'aide à la diffusion (logiciels push, outils de création de newsletters, outils pour dossiers documentaires, portails, etc.).

Aucune solution informatique ne permet l'automatisation complète de la veille, et certaines technologies sont plus ou moins bien adaptées à telle étape ou tel type de documents (exemple : l'analyse de documents structurés). On utilise assez fréquemment l'association de plusieurs "briques technologiques" pour mener à bien un process de veille automatisé.

2.4.3 La veille "manuelle" (sans l'utilisation des agents)

* **Repérer les nouveaux sites dans un domaine :**

La meilleur méthode : bouche à oreille, abonnement à des listes de diffusion, à des e-zines et newsletters.

Les services "Nouveautés" des moteurs sont trop généralistes pour être efficaces. Si votre veille s'exerce sur un secteur géographique donné, n'oubliez pas les annuaires et moteurs géographiques.

* **Suivre l'actualité :**

Cela est possible grâce aux services de diffusion personnalisée, (ex techniques de « push », via un agrégateur de fils rss (type Google Reader...)

* **S'abonner aux périodiques électroniques des sites portails importants**

Y sont indiqués le plus souvent non seulement les nouveautés du site, mais aussi du secteur concerné.

* **Quelques pistes en veille technologique :**

* Utiliser les newsgroups et les listes de diffusion scientifiques (généralement de bonne qualité)

* Utiliser les fonctions d'alerte des grands fournisseurs d'information : http://www.tictocs.ac.uk/ (devenu http://www.journaltocs.ac.uk/), scirus (diffusion de tables des matières sur profils via e-mail), ou le TOC Alert de Publist.com, Inist (veille documentaire)… Equivalent en France : le réseau Mir@bel http://www.reseau-mirabel.info/

Application de la technique « Related: » de google

http://www.google.fr/search?q=related%3Atictocs.ac.uk

http://www.google.fr/search?q=related%3Ascirus.com

* Accès plus facile et moins cher à des bases de données, par exemple de brevets (INPI www.inpi.fr)

* **Quelques pistes en veille concurrentielle ou marketing:**

* Suivre les sites web de sociétés avec un agent d'alerte comme google alert, wysigot, websitewatcher, changedetection... ou un aspirateur de sites,… ou manuellement

* Utiliser les services Push type PRLINE ou Companynews (www.prline.com)

* Utiliser les newsgroups en faisant des recherches par noms de sociétés (attention à la fiabilité de l'information !) Cela peut être toutefois un bon moyen de détecter les rumeurs et les bruits qui circulent.

2.5 Choix des mots-clés

1. Comment choisir ses mots-clés ? 201
Quand? 201
Quel type de mots-clés ? 201
Dans quelles langues ? 201
Un ou plusieurs mots-clés ? 201
Pour ou contre le SAUF? 202
Majuscules, minuscules, accents ? 202
Troncatures ? 202
Et les synonymes? 203
Mots-clés ou tags ? 205

Quand ? La sélection des mots-clés s'effectue après le choix d'une stratégie de recherche. En effet, le choix sera fondamentalement différent si l'on cherche un portail thématique, ou une source susceptible de fournir l'information ou l'information précise immédiatement. Pour simplifier, disons que dans le premier cas, les mots-clés seront "le plus large possible", dans le second cas, ils seront "le plus précis possible".

Un ou plusieurs ? On procédera par étape pour affiner éventuellement sa recherche à l'aide de plusieurs mots-clés. Si le nombre de résultats est faible avec un seul mot-clé précis (exemple : 100 résultats sur un moteur), inutile de préciser davantage. Donc, utiliser d'abord un seul mot clé (ou expression) quant la terminologie ou l'association terminologique est très spécifique. Sinon, travailler du plus général au plus spécifique (mais choisir les synonymes appropriés pour le terme générique : par exemple si je m'intéresse à un film qui s'appellerait "Demain, dès l'aube", on pourrait écrire 'film OR cinéma "demain, dès l'aube"').

Pour ou contre le SAUF ? On peut aussi isoler les mots-clés à exclure absolument car générateurs de bruit (opérateur SAUF ou signe -). Attention toutefois à ne pas aller trop vite, de peur de passer à côté de documents pertinents : Ainsi, si je cherche des informations sur les énergies alternatives autres que solaires, je peux être tenté d'"envoyer" au moteur une équation du type +"énergies alternatives" –solaires. Mais je n'aurai pas alors les ressources qui abordent successivement **toutes** les énergies alternatives. C'est pourquoi il est parfois plus judicieux de repérer une notion discriminante de son sujet de recherche plutôt que de d'utiliser sans réflexion le SAUF.

Majuscules, minuscules, accents ? De façon générale, les moteurs sont insensibles à la casse des caractères et retourneront le même nombre de résultats pour python, PYTHON, ou Python. La situation est plus contrastée en ce qui concerne les accents : si MSN, Exalead, Voila et en théorie Google traitent de manière identique les mots accentués ou non (l'expérience montre que c'est loin d'être toujours évident sur Google), Yahoo par exemple procède différemment : ils ne retournent pour un mot-clé accentué que les mots contenant l'accent, mais pour une

requête non accentuée, ils retournent les mots avec ou sans accent. Bref, il convient de faire attention à l'utilisation des accents avant d'utiliser un moteur.

Troncatures ? La troncature permet de remplacer plusieurs caractères sur la fin des mots, mais cette possibilité devient fort rare sur le web : les trois grands moteurs Google, Yahoo et MSN ne proposent pas cette option, et notamment, un mot-clé indiqué au singulier sera traité comme tel. Il est donc important de prévoir au moins l'alternative du mot au pluriel, sous peine d'occulter de nombreux résultats. pertinents. En revanche, le challenger Exalead supporte la troncature avec le caractère * Sur le Guide du web de Voila, une recherche au singulier ou au pluriel donne en revanche les mêmes résultats (faux sur le moteur Voila).

Ordre des mots ? Il peut avoir de l'importance selon les moteurs, non pas bien entendu pour le nombre de réponses, mais pour le classement des résultats : c'est par exemple vrai pour Google ou pour Voilà.

Et les synonymes ? Il est important d'explorer la terminologie du domaine de recherche, pour repérer les synonymes (très rares sont les moteurs travaillant sur les concepts). De façon générale, les premiers documents intéressants récupérés permettent de valider, compléter ou revoir ses mots-clés.

Astuces pour identifier des synonymes et/ou mots associés

* Utiliser un dictionnaire de synonymes tel celui du laboratoire de linguistique du CNRS pour les termes en français pour le français et l'anglais http://dico.isc.cnrs.fr/ (ou http://www.crisco.unicaen.fr)
* Utiliser un thésaurus de son domaine (en ligne gratuit, ou acheté comme par exemple celui de la base INSPEC (http://www.theiet.org/resources/library/index.cfm) ou la liste de Sylvie Dalbin sur http://www.dmoz.org/World/Fran%C3%A7ais/R%C3%A9f%C3%A9rences/Th%C3%A9saurus/

* Utiliser un moteur de recherche travaillant à partir de dictionnaires, encyclopédies, thesaurus, tel pour les termes en anglais FreeDictionary http://www.thefreedictionary.com/ ou le canadien http://www.granddictionnaire.com . Voir aussi www.thesaurus.com.
* Faire une recherche sur une base de données bibliographique du domaine dans lequel se situe le sujet, utilisant une indexation manuelle (dewey, autre plus spécialisée avec un thésaurus par exemple). Repérer alors comment sont indexés quelques documents pertinents, quelle est la terminologie retenue.
* Utiliser l'option define: sur Google (aujourd'hui disponible en français). On a aussi l'option Google Suggest

Utiliser les générateurs de mots clés des grands moteurs publicitaires. Attention, on travaille alors à partir des requêtes des utilisateurs beaucoup plus qu'à partir de la terminologie des documents traitant du sujet (même si parfois on a aussi l'indication des mots-clés accompagnant souvent le mot demandé dans les pages web). Ces outils servent en général à mieux référencer un site web. Ils peuvent néanmoins donner des idées (pour compléter cette liste, voir la rubrique consacrée au sujet par le site Abondance : http://www.abondance.com/ressources/generateur-mot-cle.html

- Générateur de mots-clés du programme Google Adwords : *https://**adwords**.google.com/o/KeywordTool*
- Outiref (www.outiref.com). Voir aussi WebRankinfo http://www.webrankinfo.com/outils/semantique.php

* Utiliser pour l'anglais le méta-moteur Surfwax (www.surfwax.com) en cliquant sur la petite flèche suivant la ligne "focus:mot-clé choisi" au dessus des résultats à gauche : Notons que Surfwax a mis en ligne (mars 2005) WikiWax, qui combine l'outil de suggestion de terme de Surfwax et l'encyclopédie collaborative en ligne Wikipedia. http://www.wikiwax.com/.
* Explorer les balises méta (keywords) de quelques documents pertinents

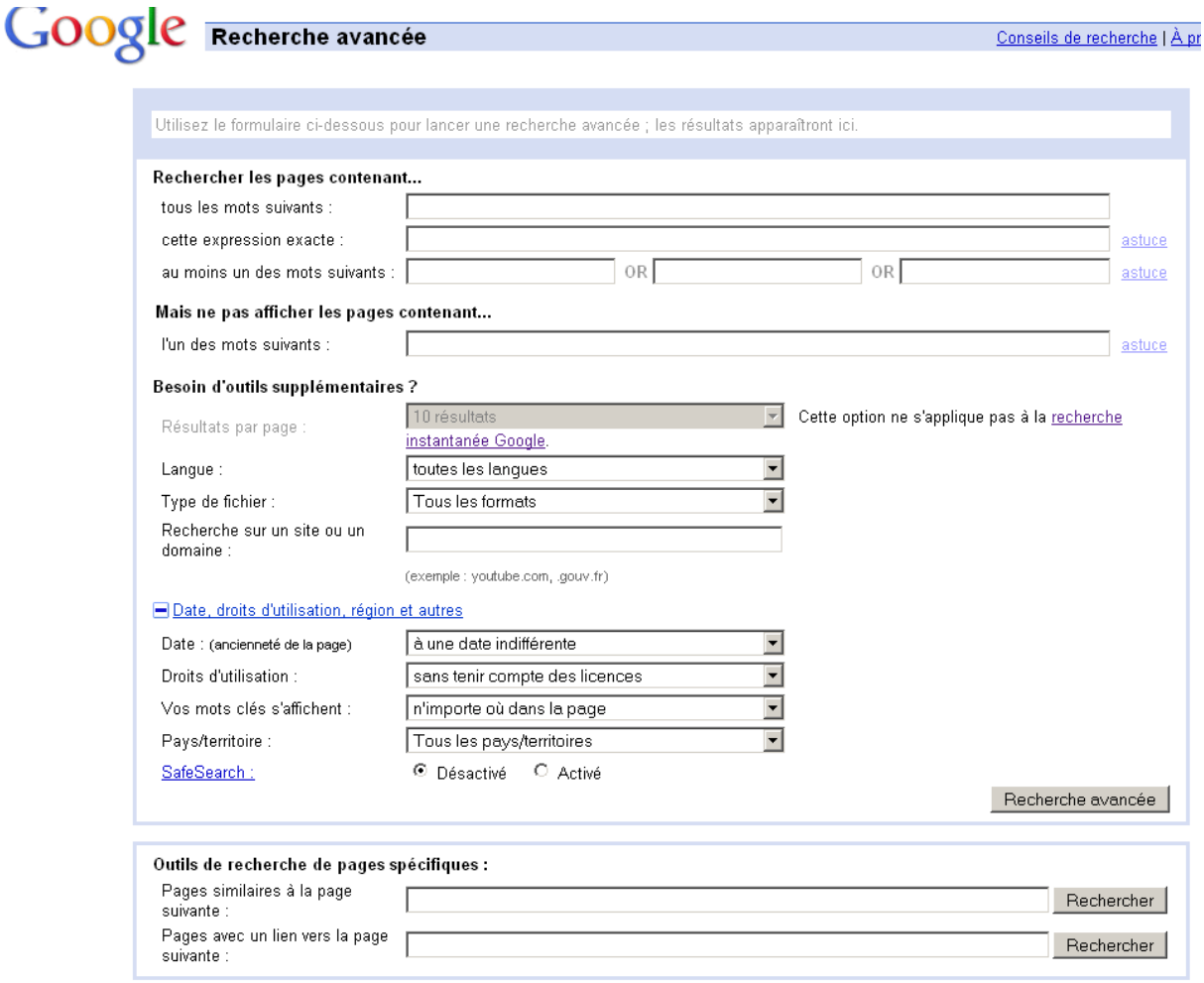
Pour passer du français à l'anglais, utiliser à partir d'une catégorie donnée, le "passage direct" de yahoo.fr à yahoo.com : "Poursuite de la recherche sur Yahoo US". On peut aussi faire une recherche moteur avec un mot-clé large en français et en anglais et un mot-clé "profond" en français seulement : wine vin soutirage peut me donner des infos terminologiques sur le soutirage en anglais.

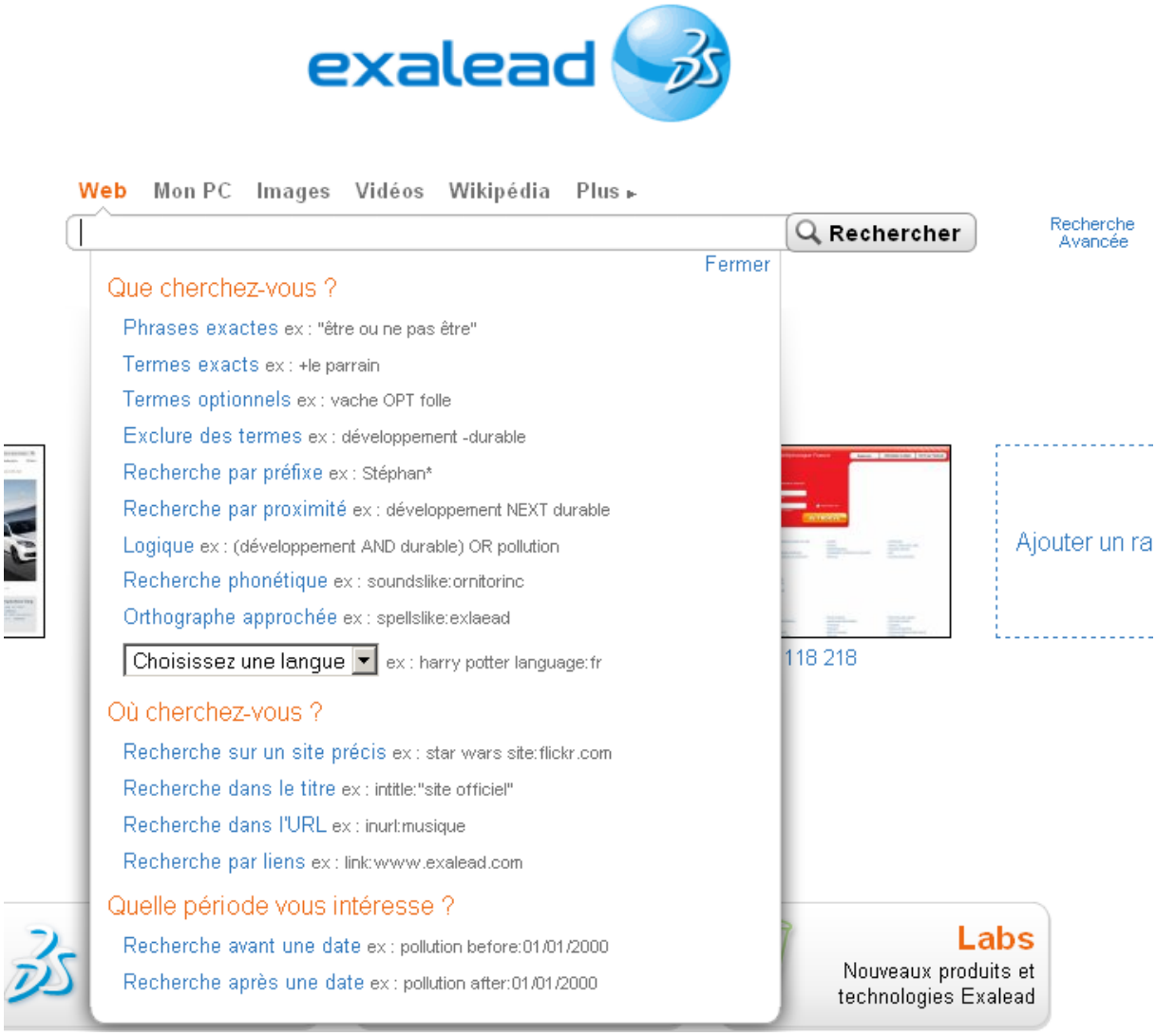
2.5.1 Eurovoc

http://eurovoc.europa.eu/drupal/?q=fr

2.6 Opérateurs avancés de recherche

2. Quels sont les opérateurs de recherche indispensables ? 206
1 Les guillemets « »… pour manipuler les expressions 206
2 Le +… pour imposer un mot ; le – … pour l'exclure 206
3 intitle:… pour chercher dans le titre 207
4 site:… pour cibler un domaine 207
5 filetype:… pour trouver les documents directement
dans le bon format 207




exalead
Web
Mon PC
Images
Vidéos
Wikipédia
Plus ▸
Rechercher
Recherche Avancée
Fermer
Que cherchez-vous ?
Phrases exactes ex : "être ou ne pas être"
Termes exacts ex : +le parrain
Termes optionnels ex : vache OPT folle
Exclure des termes ex : développement -durable
Recherche par préfixe ex : Stéphan*
Recherche par proximité ex : développement NEXT durable
Logique ex : (développement AND durable) OR pollution
Recherche phonétique ex : soundslike:ornitorinc
Orthographe approchée ex : spellslike:exlaead
Choisissez une langue
ex : harry potter language:fr
Où cherchez-vous ?
Recherche sur un site précis ex : star wars site:flickr.com
Recherche dans le titre ex : intitle:"site officiel"
Recherche dans l'URL ex : inurl:musique
Recherche par liens ex : link:www.exalead.com
Quelle période vous intéresse ?
Recherche avant une date ex : pollution before:01/01/2000
Recherche après une date ex : pollution after:01/01/2000
118 218
Ajouter un ra
Labs
Nouveaux produits et technologies Exalead

http://onlinehelp.microsoft.com/fr-FR/bing/ff808535.aspx

Bing Help Home >

Imprimer | Courrier é

Aide de Bing

Découvrir la Bing

Optimiser la recherche

Search tips and techniques

Bing FAQ

Fonctionnalités utiles de la Bing

Obtenir des suggestions de recherche

Afficher l'historique de vos recherches

Trouver des réponses instantanées

Découvrir la page d'accueil enrichie

Décidez avec l'aide de vos amis

Traduire les résultats de la recherche

Trouver ce dont vous avez besoin

Retrouver les photos de vos artistes préférés

Rechercher et regarder des vidéos

Rechercher des produits en ligne

Voir les avis des utilisateurs d'un coup d'œil.

Obtenir les dernières nouvelles

Trouver les prévisions météorologiques

Rechercher des sites Web cinématographiques, de produits et de sociétés

Find flights with low airfares

Get alerted when airfares change

Find flight deals

Find hotel deals

Compare travel dates and destinations

Find out what's new in Bing

Limiter vos résultats

Utiliser la recherche avancée

Options de recherche avancée

Recherche avancée par mots clés

Résoudre un problème

La recherche n'a pas renvoyé les résultats appropriés

Peu de résultats, voire aucun

Trouver des informations sur des messages d'erreur sur le Web

Why some results have been removed

Référence

Notations mathématiques à utiliser avec les Réponses mathématiques

Codes de pays, de région et de langue

En savoir plus sur la sécurité et la confidentialité

À propos de la déclaration de confidentialité et de la stratégie de sécurité de Microsoft

Méthodes utilisées par Bing pour proposer des résultats de recherche

Influence des annonces sur les résultats de recherche Bing

Bloquer des sites Web explicites

Signaler un problème concernant un résultat

Soumettre un problème de marque

2.7 Astuces pour identifier rapidement des sources d'information, des experts, un type de contenu spécifique, etc.

(Voir chapitre : Evaluer une information sur Internet)

- Usage des commandes google : related: , link: (popularité, similarité, recherche de « retroliens »,...)
- Présence dans les forums, mentions dans les sites sociaux (bon ou « bad-buzz »)
  - Ebuzzing (ex-wikio)
  - Twitter
  - Newsgroups ou liste de discussion...
- Respect des standards (d'où connaissance des référentiels du W3C, WAI, RGI, RGS, RGAA....)
- Nom de domaine (connaissance et respect des normes d'URL, occupation des espaces de noms, taille et graphie de l'URL, choix de l'extension du nom de domaine...
- Présence de meta-tag dans le source des pages
- Bons usages en matière d'ergonomie en général (couleurs, taille d'écran, langue, orthographe...)
- Respect de la législation et réglementation française et européenne en matière de diffusion d'information sur internet : mentions légales, CNIL, DAVDSI, LCEN,...
- Connaissance et liens vers les sites « pivots » (incontournables dans la matière : sites officiels, gouvernementaux,...)
- Plus généralement : E-réputation : présence sur le web, sur Wikipedia, ...

3 *Apport des outils et pratiques du web 2.0 : en quoi sont-ils créateurs de valeur ?*

3.1 Archives ouvertes

Les ressources géantes de l'OAI-PMH

Le concept d'archives libres et de dépôts institutionnels

http://repositories.webometrics.info/

July 2011

Top Repositories

First | Previous | Next | Last | Repositories 1 to 50 of 1222

WORLD RANK	REPOSITORY	COUNTRY	POSITION SIZE	VISIBILITY	RICH FILES	SCHOLAR
1	Social Science Research Network		6	2	1	4
2	Arxiv.org e-Print Archive		4	3	10	3
3	CiteSeerX		1	1	1,070	2
4	Research Papers in Economics		2	5	174	5
5	Smithsonian/NASA Astrophysics Data System		3	4	894	1
6	CERN Document Server		7	17	2	8
7	National Taiwan University Repository		16	7	27	7
8	Kyoto University Research Information Repository		13	9	3	65
9	HAL Sciences de l'Homme et de la Société		65	13	38	30
10	Munich Personal Repec Archive		56	22	17	26
11	University of California eScholarship Repository		18	47	8	15
12	HAL Institut National de Recherche en Informatique et en Automatique Archive Ouverte		52	27	23	25

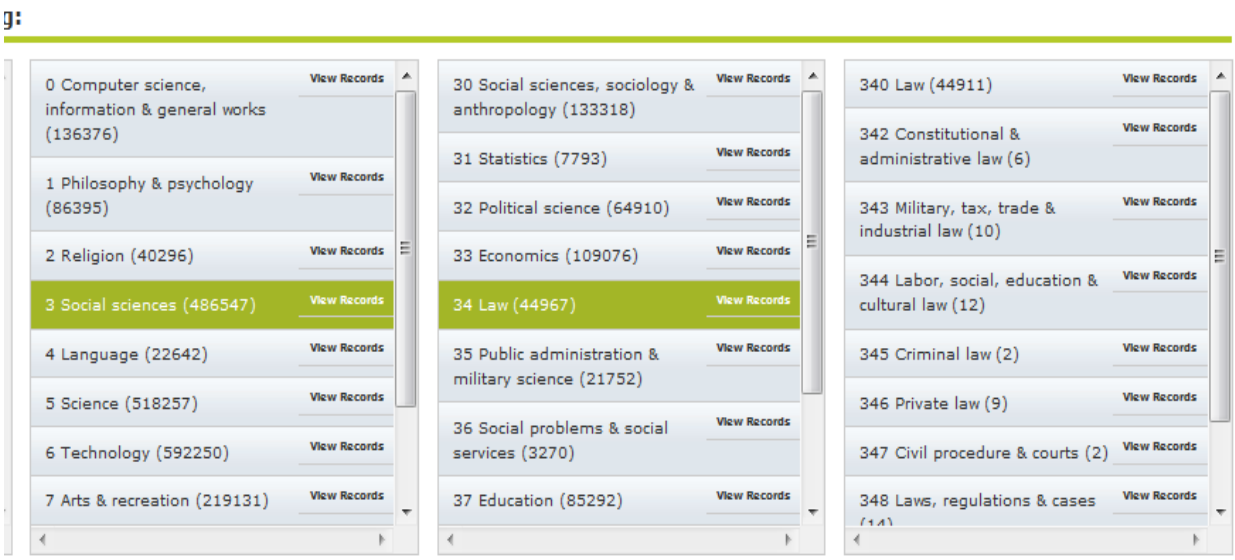
Alternative à Google Scholar : Initiative Base Search de Bielefeld : (base-search.net)
http://www.base-search.net/about/en/index.php

BASE is one of the world's most voluminous search engines especially for academic open access web resources. BASE is operated by Bielefeld University Library.

As the open access movement grows and prospers, more and more repository servers come into being which use the "Open Archives Initiative Protocol for Metadata Harvesting" (OAI-PMH) for providing their contents. BASE collects, normalises, and indexes these data. The Index more than 30 million documents from more than 2,000 sources. You can access the full texts of about 75% of the indexed documents. The Index is continuously enhanced by integrating further OAI sources as well as local sources. Our OAI-PHM Blog communicates information related to harvesting and aggregating activities performed for BASE.
BASE is a registered OAI service provider and contributed to the European project "Digital Repository Infrastructure Vision for European Research" (DRIVER). Database managers can integrate the BASE index into your own local infrastructure (e.g. meta search engines, library catalogues) via an interface.

In comparison to commercial search engines, BASE is charcterised by the following features:

- Intellectually selected resources
- Only document servers that comply with the specific requirements of academic quality and relevance are included
- A data resources inventory provides transparency in the searches
- Discloses web resources of the "Deep Web", which are ignored by commercial search engines or get lost in the vast quantity of hits.
- The display of search results includes precise bibliographic data
- Several options for sorting the result list
- "Refine your search result" options (by author, subject, DDC, year of publication, collection, language and document type)
- Browsing by DDC (Dewey Decimal Classification) and document type.

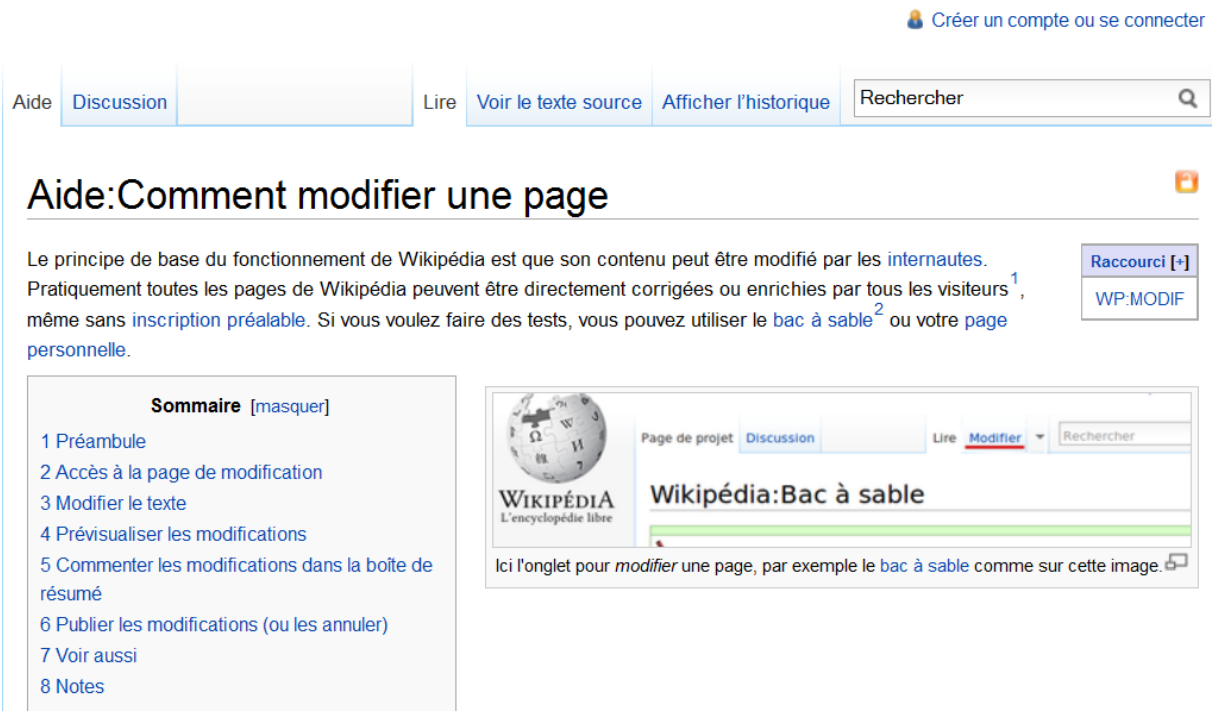


3.2 Blogs, wikis?

3.2.1 Créer un blog

http://www.blogger.com

3.2.2 Modifier une page wiki



Application : survol des techniques wiki et notamment de la wikipedia. Apprentissage de la modification de pages et de l'analyse des historiques.

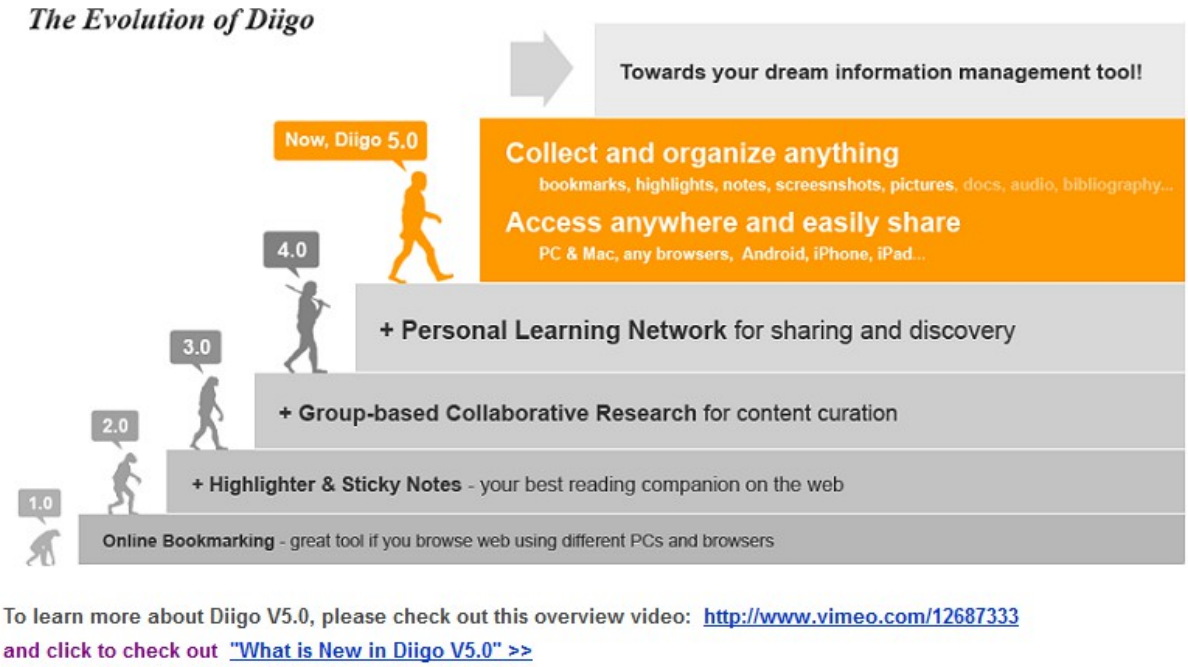
3.3 Tags et folksonomie

3.3.1 Gérer un espace de signets partagés et taguer : diigo

http://blogs.crdp-limousin.fr/stage-veille/2011/05/17/didacticiels-diigo/
Didacticiels DIIGO, Didier Pouzaud, sur le BiblioLab de la BNF

http://bibliolab.fr/cms/content/comment-fonctionne-diigo

Le Bibliolab est une plateforme de formation, d'expérimentation et d'information autour des TIC et du numérique en bibliothèque.



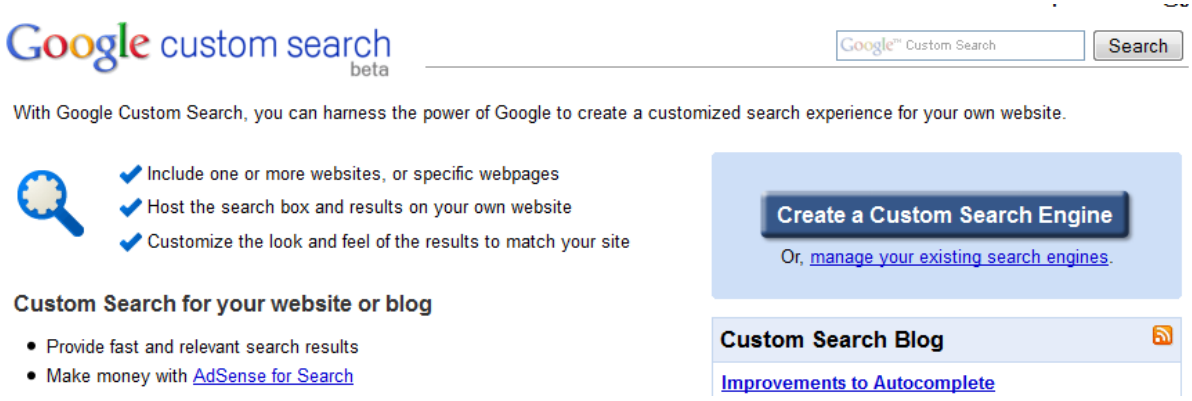
http://cursus.edu/dossiers-articles/articles/9750/diigo-mode-emploi-pour-debutants/

3.3.2 Twitter et l'usage des Hashtags

Usage d'un compte twitter
Lecture de Time Line
Reprise de mini messages
Recherche et alerte sur #hashtag (ou « mot-dièse »)

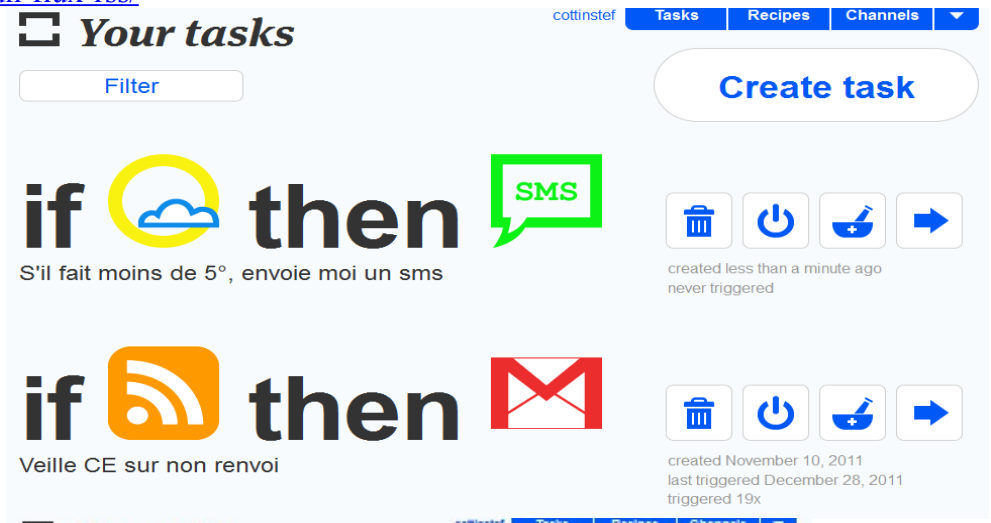
3.4 Moteurs personnalisables

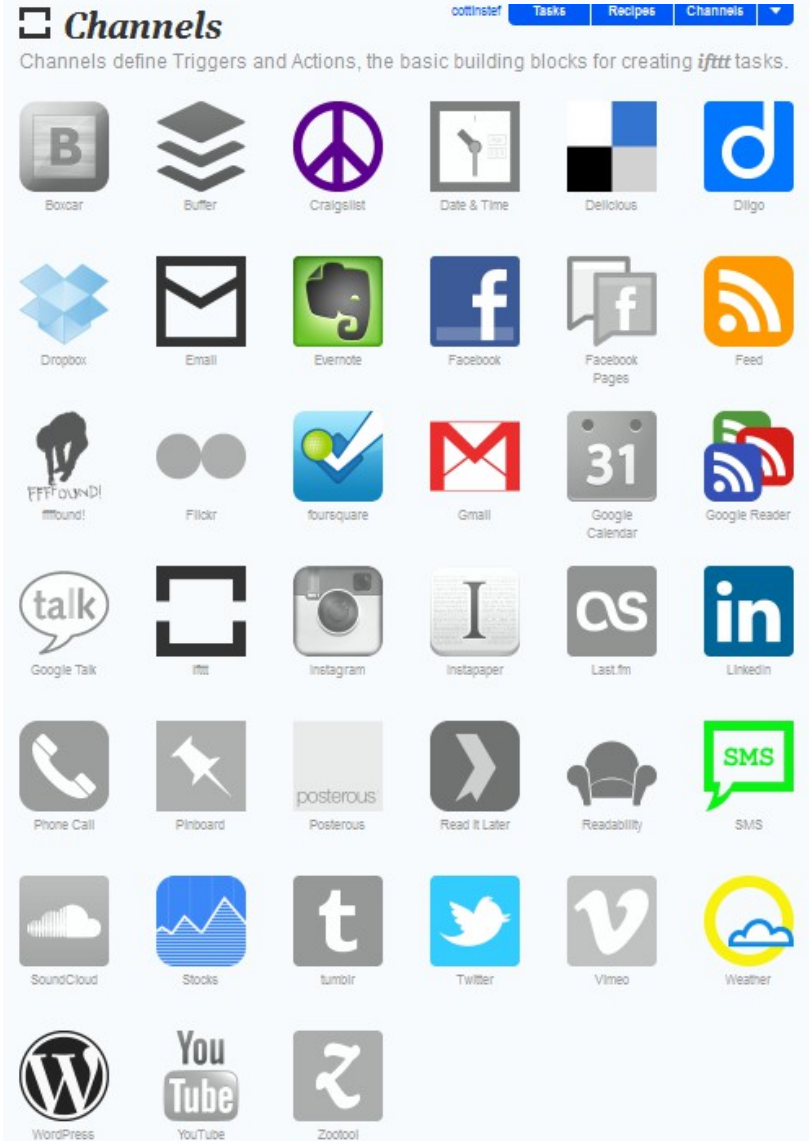
Monter un Google Custom Search Engine



3.5 Partage de signets/présentations/images/vidéos?

Les outils de mashup type Ifttt.com
http://donneesjuridiques.wordpress.com/2012/10/22/astuces-ifttt-1-creer-une-alerte-e-mail-a-partir-dun-flux-rss/





4 *Evaluer l'information sur Internet*

4.1 Quelques questions clés à se poser

Vérifier ses sources

Web bibliographie :

Information Quality Resources on the Internet, Marcus P. Zillman, Published on December 2, 2011 sur LLRX http://www.llrx.com/features/informationqualityresources.htm (Legal Librarians Ressources eXchange)

4.2 Comment évaluer un site web ?

L'évaluation de l'information sur Internet devient un enjeu important pour les professionnels. Il s'agit d'un acte d'expertise pour estimer la qualité des différentes ressources disponibles :le portail, le site web, la page web, l'article sur la page, la base de donnée accessible depuis la page, mais aussi le forum, la liste de discussion, le message posté sur une liste ou un forum, etc.

Les critères d'évaluation

Différentes catégories de critères sont à prendre en compte, sachant qu'il convient de croiser une évaluation de la source avec une évaluation du contenu :

* Crédibilité : Organisation émettrice, type d'émetteur, auteurs des documents, source de financement ou sponsoring, webmaster, cibles et objectifs du site, type d'accès, etc.
* Fraîcheur : Date de création et de mise à jour
* Exhaustivité et l'exactitude : Type de document, citations des sources, bibliographie, contextualisation de l'information, qualité de la langue, etc.
* Adéquation : pertinence et utilité par rapport à la recherche ou à la veille menées.
* Ergonomie : arborescence, navigation, orientation, frames, etc.
* Design : présentation visuelle, conception graphique.

Les grilles d'évaluation existantes

La plus aboutie sur le Web (mais très lourde) dans le domaine de l'information santé http://www.chu-rouen.fr/netscoring

Voir aussi

Université Laval *www.fse.ulaval.ca/fac/href/grille/grille.gif*

Il est intéressant de consulter le cours en ligne "L'évaluation de l'information sur Internet" et les textes déposés sur les archives institutionnelles à l'adresse http://urfistreseau.wordpress.com/les-intervenants/alexandre-serres/ , élaboré par Alexandre Serres, responsable URFIST Bretagne

Astuces pour l'évaluation des pages en cours de navigation

* *Chercher des informations sur l'éditeur sur le site*. En cas de difficulté, chercher le copyright en bas de page. On peut aussi repérer sur le plan du site la page Contact qui va fournir un email. Voir alors la seconde partie de l'adresse mail (après le @) qui peut renvoyer à un domaine particulier que l'on cherchera alors sur le web.
* *Chercher des informations sur la société indiquée*. On utilisera alors des bases de données d'informations sur les sociétés (R5CS, organismes de régulation boursiers).
* *Pour rechercher le propriétaire d'un nom de domaine* (noms des responsables techniques et administratifs). Attention, les informations sont loin d'être toujours mises à jour, donc il y a des risques d'erreur, et parfois besoin de recoupements.
    - Pour les noms de domaine se terminant par un ".fr" on utilisera le moteur proposé par l'AFNIC, centre d'information et de gestion des noms de domaine pour la France (et pour l'île de la Réunion .re) : www.afnic.fr
    - Pour les noms de domaine "gTLD" (generic Top Level domains), c'est à dire les .com, .net, .org, et plus récemment les .biz et les .info, c'est plus difficile car les bases de données ne sont plus unifiées (auparavant, la base Whois gérée par l'Internic). On utilisera donc un méta-moteur comme Betterwhois, qui permet d'interroger les bases des "régistrants" (prestataires assurant la gestion administrative et technique du nom de domaine) les plus importants : www.betterwhois.com.
    - Pour les autres noms de domaine par pays, on peut passer par un service générique http://www.generic-nic.net/dyn/whois, ou bien chercher préalablement l'organisme national pays par pays sur Yahoo : http://dir.yahoo.com/computers_and_internet/internet/domain_name_registration/top_level_domains__tlds_/registry_operators/International_Country_Codes/
* *Pour trouver des informations générales sur la page*, on peut utiliser le moteur Alexa www.alexa.com, propriété de Amazon.com. On obtient les coordonnées du "régistrant", mais aussi des statistiques sur le trafic du site, des témoignages d'internautes, le temps de chargement de la page, le nombre de liens vers cette page, etc. De plus, des sites/pages "similaires" sont proposés.
* *Utiliser également le "URL info" de Fagan Finder :* http://www.faganfinder.com/urlinfo
* *Ne pas oublier non plus de faire des recherches sur le web* en prenant le nom du site comme mot-clé, et avec la fonction link : (recherche par popularité : qui a un lien sur cette page).
* *On peut aussi utiliser l'interface de recherche développé par un journaliste Jean-Marc Manack pour se simplifier la vie dans la validation des informations :* Plus de 200 outils classés par rubriques (moteurs de recherche, administratif – URL, dictionnaires, référence, actualités, blogs, etc.) sont disponibles à partir d'un seul formulaire, les résultats apparaissant dans la partie gauche de la page. La différence avec un méta-moteur classique, est que l'on peut mettre soit un mot-clé, soit une url. (utilisable aussi en mode "sidebar" dans le navigateur : http://manhack.net

4.3 Quelques outils pratiques

4.3.1 WHOIS

(voir évaluation)

4.3.2 Où trouver des archives du web ?

Rien n'est exhaustif dans le monde du web, mais le service proposé par l'association The Internet Archive (qui reçoit des donations et soutiens de différents acteurs, dont Alexa) est très impressionnant : on peut ainsi visualiser un site tel qu'il était à différentes dates depuis 1996, et même suivre des liens sur ces archives.

The way back machine : www.archive.org

Depuis la fin 2003, un service en beta permettait d'aller beaucoup plus loin, en permettant une recherche plein texte, par date, sur plus de 11 milliards de pages archivées. Différentes fonctionnalités étaient accessibles à partir des résultats des sites répondant le mieux à la recherche : graphique permettant de voir la fréquence d'apparition du mot-clé sur la période, thèmes traités par le site, concepts proches, etc. : http://recall.archive.org

4.3.3 Comment trouver des bookmarklets ?

Les bookmarklets sont des programmes contenus dans des liens, c'est à dire des éléments de code java qui se mettent dans les favoris comme des URL classiques, mais qui déclenchent quand on les appelle une action particulières. Ils déclenchent souvent ouverture de fenêtre pop-up (ce qui pose d'ailleurs un problème quand on utilise un "anti pop-up" : obtenir le premier résultat du moteur Google directement, faire un lien direct vers un paragraphe de page html, traduire, éditer les urls présents sur une page à la fin de celle-ci, intégrer un nouveau bookmark si l'on est sur un service en ligne de gestion de favoris, etc.

Pour en trouver, et pour démarrer votre recherche : - http://www.outilsfroids.net/texts/OutilsBookmarklets
www.bookmarklets.com

4.3.4 Comment gérer les problèmes fréquents avec les outils ?

* **Erreurs 404, liens non valables** : remonter dans la hiérarchie du site. Si l'adresse de l'host est bonne, revenir à cette adresse et "tatonner" à l'intérieur du site pour retrouver la page cherchée et sa nouvelle URL. On peut aussi utiliser le lien "cached" sur Google ou les archives de Alexa.

* **Signification des principaux messages d'erreurs** :

Erreur	Message	Signification
400	Bad Request	Erreur dans l'adresse
401	Access Denied	La consultation nécessite un nom d'utilisateur et un mot de passe
403	Forbidden	L'accès est réservé et vous n'avez pas les privilèges correspondants
404	Not found	La page correspondant à cette URL n'a pas été trouvée sur le serveur
500	Internal	Problème de serveur. Contacter l'administrateur du site
503	Read time out	Le temps alloué à la connexion est écoulé

* **Réponses hors sujet** : reformuler sa question, rajouter des mots clés…
* **La page proposée ne contient pas votre terme de recherche** .

Il peut y avoir plusieurs explications, mais la plus vraisemblable est que ce mot se trouvait dans la page lorsque celle-ci a été sauvegardée par le robot du moteur. Puis elle a été modifiée et le mot a disparu de la page. Mais par contre il est resté dans l'index de la base de données. Il se peut aussi que votre terme apparaisse dans un formulaire déroulant, ou enfin en méta-données.

Une solution pour être certain d'obtenir des résultats contenant les mots-clés de votre question consiste à utiliser un méta-moteur "off-line" avec la fonction "raffiner" ou "filtrer".

* **Non élimination des doublons** : les moteurs utilisent maintenant à peu près tous les techniques de clustering pour la présentation des résultats (une réponse = un site et non une réponse = une page) ou le proposent en option. Mais cela n'empêche pas toujours les doublons.
* **Problème d'accès à de l'information très récente** : attention, un moteur peut mettre plusieurs jours ou mêmes semaines avant d'indexer un nouveau site… Voir du côté des serveurs d'actualité, par exemple.

4.3.5 Peut-on circuler de façon anonyme sur le web ?

On le sait, la navigation sur le web laisse des traces (voir notamment à ce sujet le site de la CNIL www.cnil.fr). Il existe néanmoins des services permettant de masquer les adresse IP d'origine et d'empêcher les cookies et autres techniques de marquage de fonctionner, c'est à dire de garantir une meilleure confidentialité de surf sur internet

Anonymiser http://www.anonymizer.com/ (payant)

Voir le TOR Project Anonymity online https://www.torproject.org/

Voir enfin le portail Stay Invisible qui propose définitions, actualités, tests, un forum de discussion sur le sujet ainsi qu'une liste d'outils **: http:**//www.stayinvisible.com

4.3.6 Peut-on effectuer des traductions de textes sur le web ?

Des outils gratuits sont disponibles en ligne pour traduire des textes, voire des pages web. Les résultats sont certes souvent discutables, mais pour une première approche, ces technologies peuvent être d'une aide réelle à la recherche.

Sur Voila (technologie Systran) http://tr.voila.fr

Sur Google (technologie Systran) http://translate.google.com/

Sur Alta Vista (technologie Systran) http://babelfish.altavista.com/ / http://fr.babelfish.yahoo.com/

Sur Reverso (technologie Reverso) http://www.reverso.net

5 *Autoformation*

5.1 Veille sur l'actualité des outils de recherche d'information

http://motre.ch/ , Jérôme Charron ,Emilie Ogez ,Frederic Martinet
http://www.abondance.com/ L'actualité des moteurs et du référencement... Olivier Andrieu
http://www.les-infostrateges.com/ publie en permanence des actualités, des articles et des dossiers de fond qui rendent compte de l'expertise des auteurs en stratégies informationnelles. (spécialisés en droit de l'information et en "e-reputation")

http://sapristi-docinsa.insa-lyon.fr/guides-similaires (Guide Sapristi, Insa de Lyon)
Les guides francophones

- L'URFIST de Lyon propose des documents très intéressants dont :
- Des informations variées sur Internet
- Un guide des bases de données gratuites sur le web : DADI
- Un guide consacré aux sociétés et au marché des bases de données sur le web : SINBAD
- Des documents pédagogiques relatifs à l'actualité du web
- L'URFIST de Rennes propose de nombreuses pages dont des supports de formation indispensables à la recherche sur Internet. : http://www.sites.univ-rennes2.fr/urfist/ressources
- L'URFIST de Paris propose CERISE : Conseil aux Etudiants pour une Recherche d'Information Spécialisée et Efficace.
- ABCdoc est un guide méthodologique de recherche et de traitement de l'information scientifique et technique produit par la SUP (Structure Universitaire de Pédagogie) de l'Université Paul Sabatier (Toulouse 3)
- Form@doct est un guide destiné aux doctorants complément indispensable à Sapristi ! mais pouvant être utilisé par des d'étudiants d'autres niveaux . Il est produit depuis 2010 par l'UEB (Université Européenne de Bretagne)
- Méthodoc est un guide de méthodologie documentaire et de sciences de l'information. Il est destiné aux étudiants du supérieur. Il est co-produit par le SCD de l'Université Rennes 2 et par l'URFIST de Rennes.
- On peut aussi rajouter à cette liste non-exhaustive les pages signets de la BNF, de la BPI et du CERIMES

Les guides anglophones

- A short and easy search tutorial de Pandia
- Evaluating Quality on the Net . mis en ligne en 1995 depuis l'université de Harvard, ce site propose une approche pour l'évaluation de l'information sur internet et indique des outils pour rechercher cette information.

5.2 Veille sur les producteurs de sources d'informations

GFII http://www.gfii.fr/fr/

http://conferences.isko-france.asso.fr/fr/index.htm

Internet Actu et FING http://www.internetactu.net/

Le Journal du Net : e-Business, Informatique, Economie et … *www.journaldunet.com*
Analyses, tendances, interviews : tous les jours, le Journal du Net vous propose le meilleur de l'actualité Internet et e-business.

affordance.info *affordance.typepad.com*
Notes, liens et réflexions en rapport avec les sciences de l'information et la gestion des connaissances, par Olivier Ertzscheid.

Business et Solutions IT - Toute l'actualité de l'internet et du marché … *www.zdnet.fr › News* -
ZDNet.fr, le site d'information pour les décideurs et les utilisateurs professionnels IT en France. Retrouvez l'actualité de l'internet et du marché …

01net informatique high-tech : actu, produits, téléchargement … *www.01net.com/* -
Toute l'actualité informatique et high tech : emploi, test de produits high-tech et jeux vidéos, astuces et logiciels à télécharger. Découvrez les dossiers et conseils ...