4 La recherche sur l'Internet

La recherche d'informations sur l'Internet pose au moins trois problèmes:

1. La quantité d'informations disponibles est colossale
2. On ne sait pas sur quel site aller pour trouver une information intéressante
3. N'importe qui peut venir déposer des informations sur l'Internet; même des informations fausses

Dans une bibliothèque classique, on peut faire appel au bibliothécaire. Mais personne n'est chargé du rôle de bibliothécaire sur l'Internet.

1 Les outils de recherche sur l'internet

Comment faire une recherche sur Internet ?

Meta-moteurs, annuaires ou moteurs de recherche, tous concourent à vous aider à trouver ce que vous recherchez sur le web. Pourtant leur mode de fonctionnement, leurs avantages et inconvénients diffèrent fortement.

Les moteurs de recherche (Altavista, Lycos, Google, …)

- **Pour quelle recherche ?**

  Sur un mot clé précis, une expression ou une citation exacte.

- **Méthode de recherche**

  Régulièrement, des robots explorent le web en sautant de lien en lien et "attrapent" des mots-clés, que cela soit dans le texte ou dans les "metatags", ces mots-clés introduits par l'auteur de la page dans le code HTML. Ils indexent ensuite les pages selon leur degré de pertinence (ex : Google les classe par rapport au nombre de liens pointant sur la page).

- **Avantages**

  C'est avec ce type d'outil que vous pourrez explorer le plus largement le web et obtenir le plus grand nombre de pages.

- **Inconvénients**

  Les résultats que vous obtiendrez seront toujours anciens : en effet, il faut du temps pour que les robots explorent le web, puis que les pages soient indexées dans le moteur. En outre, si votre demande n'est pas extrêmement précise, vous risquez de vous voir proposer des milliers de réponses.

- **Les méta-moteurs (Metacrawler…)**
- **Pour quelle recherche ?**

Ce genre d'outil est utile pour ceux qui veulent un rapide tour d'horizon général d'un sujet ou d'un mot-clé.

- **Méthode de recherche**

Ces méta-moteurs n'ont pas de base de sites propre, ils explorent plusieurs moteurs de recherche à la fois. Certains vous procureront une liste unique de résultats, d'autres en procureront plusieurs selon le moteur d'origine.

- **Avantages**

Rapidité et recherche simultanée sur plusieurs moteurs donc : beaucoup de résultats.

- **Inconvénients**

Peu consultent Northern Light ou Fast Search qui possèdent les plus grandes bases de données disponibles sur le web. La quantité de réponses prime souvent sur la pertinence. Beaucoup de doublons.

- **Les annuaires (Yahoo !, Voilà, Excite…)**
- **Pour quelle recherche ?**

Pour ceux qui ne savent pas exactement ce qu'ils cherchent ou qui font des recherches très générales (par exemple : tourisme, immobilier…).

- **Méthode de recherche**

Ce sont les seuls outils gérés par des humains. Ce ne sont pas des bases de données de pages mais simplement de liens menant vers des pages. Les rédacteurs de ces annuaires explorent et sélectionnent les pages web qu'ils classent ensuite dans des catégories.

- **Avantages**

- Le côté humain garantit la pertinence des résultats.

- La navigation dans les catégories permet de préciser et de concentrer sa recherche et donne des idées de mots-clés à soumettre aux moteurs. La plupart des annuaires sont d'ailleurs associés à des moteurs qui permettent d'obtenir encore plus de résultats.

- **Inconvénients**

Le contenu des pages indexées peut changer ou les pages ne plus exister sans que les rédacteurs le sachent.

Distinguer les annuaires (portails) et les moteurs

Beaucoup d'internautes ont tendance à utiliser de la même manière les annuaires (comme Francité ou Open Directory) et les moteurs de recherche (comme Google ou Bing). Ces deux types d'outils sont pourtant de nature différente.

ANNUAIRES	MOTEURS
Indexation de sites - par des documentalistes	**Indexation de mots** - par des robots
Recherche - sur des sites - et des catégories	*Recherche* - en texte intégral - sur des pages web
Avantages - choix des informations - classement raisonné par catégories et sous-catégories	*Avantages* - plus d'exhaustivité, - mise à jour plus rapide
Inconvénients - moins d'exhaustivité, - mise à jour moins rapide	*Inconvénient* capture de pages web sans classement raisonné
A retenir L'exploration des catégories s'avère souvent plus fructueuse que celle des sites.	*A retenir* La recherche par mots clés donne de meilleurs résultats sur les moteurs

Exemples de moteurs de recherche:

Généralistes : Alta Vista <http://www.altavista.com/>, Google http://www.google.com/

Géographiques : Excite France <http://www.excite.fr/>

Spécialisés : Search4science <http://www.search4science.com/>, BioHunt <http://www.expasy.ch/BioHunt/>

Exemples d'annuaires :

Généralistes : <http://dir.yahoo.com/>

Géographiques : Woyaa <http://www.woyaa.com/>, Liege.com <http://www.liege.com/annuaire.htm>

Spécialisés : Searchengineguide <http://www.searchengineguide.com/searchengines.html>, Agaf <http://www.agaf.org/>

Distinguer les moteurs et les métamoteurs

Les métamoteurs utilisent simultanément plusieurs moteurs et annuaires.

MOTEURS	METAMOTEURS
Répétition de la requête sur chaque moteur	**1 seule requête simultanée** sur plusieurs moteurs et annuaires
Avantage une syntaxe spécifique (résultats plus précis)	*Avantage*

Inconvénient temps de réponse plus long	gain de temps *Inconvénient* pas de syntaxe commune (résultats moins précis)

Exemples de métamoteurs:

Profusion <http://www.profusion.com>,

Metacrawler http://www.metacrawler.com

- **Les sites fédérateurs et les guides - recherche large dans un domaine précis**

Aucun outil n'est exhaustif et bien souvent, il faut en utiliser plusieurs pour arriver à ses fins. Les sites fédérateurs (Gateway, Portail, passerelle thématique) et les guides thématiques sélectionnent des sources de qualité dans un domaine précis. Créés par des professionnels de l'information ou des passionnés, ces sites proposent en général un recensement complet des meilleures ressources concernant un domaine. Par ailleurs, ils recensent bien souvent des ressources appartenant au Web invisible. Un site fédérateur peut donc proposer des répertoire spécialisés, des liens vers des répertoires ou pages de liens spécialisées, des articles en texte intégral ou une bibliographie en ligne, les actualités du secteur, les événements du secteur, des accès à des base de données, des offres/demandes d'emploi, un forum, des données chiffrées, des statistiques, des synthèses concernant le secteur, une liste de périodiques spécialisés, une liste d'experts, des cours, des conseils, des informations juridiques, etc.... Un site fédérateur très actif rassemble souvent une communauté de spécialistes autour de lui et devient donc un point de référence du domaine.

Exemple:

Map History <http://www.maphistory.info/>

2 Les moteurs de recherche

Un moteur de recherche est formé d'une batterie d'ordinateurs muni de programmes et capable de trouver des informations parmi un grand nombre de documents de différents types.

Exemple: google, yahoo, Ask, bing …

3 Fonctionnement des moteurs de recherche: synthèse

3.1 Les web crawlers ou spiders

Comme leur nom l'indique, les web crawlers passent leur temps à explorer l'internet (textuellement "faire du crawl sur le web"). Ils parcourent tous les documents qu'ils trouvent en suivant les liens hypertextes.

On pourrait aussi parler de "spiders" (araignées, en anglais) qui parcourent continuellement la toile (le World WideWeb).

En "lisant" les pages web, les spiders repèrent les liens hypertextes et sautent ensuite vers les pages liées. Qu'ils lisent en repérant les liens hypertextes, et ainsi de suite.

Remarques:

- Les pages qui sont fréquemment modifiées ; les pages des journaux quotidiens, par exemple- sont parcourues plus régulièrement que d'autres.
- Les pages qui ne sont liées à aucune autre page ne sont jamais visitées

3.2 Les serveurs d'index

Les pages "lues" par les spiders sont envoyées vers une autre série d'ordinateurs: les serveurs d'index. Leur rôle est de tenir à jour un index des informations lues par les spiders.

Cet index se présente comme l'index d'un livre: à chaque mot, on fait correspondre la page où ce mot se trouve. Mais en beaucoup plus gros. Il constitue une gigantesque banque de données dans laquelle il sera possible de chercher très rapidement des informations.

4 Fonctionnement des moteurs de recherche: les requêtes

Lorsque l'on effectue une requête sur un moteur de recherche, celui-ci interroge la banque de données (les serveurs d'index) dont il dispose pour répondre à la question posée.

Définition : Une **requête** sur un moteur de recherche est une question posée à la base de données du moteur de recherche.

Les différentes phases du processus:

1. Un internaute envoie une requête sur un moteur de recherche
2. Les systèmes informatiques du moteur de recherche interrogent la banque de données des serveurs d'index
3. Les serveurs d'index renvoient une série de résultats pour la requête qui a été transmise
4. Le serveur web du moteur de recherche retourne les résultats à l'internaute

5 Analyse des résultats de recherche

(voir l'exposé)

6 Affiner le travail sur un moteur de recherche

Dans certains cas, les moteurs de recherche renvoient des résultats sans rapport avec le sujet qui nous intéresse. Comment éviter cela et mieux cibler la recherche?

➢ Bien choisir les mots-clés utilisés

- Avant de commencer toute recherche, **éteindre l'ordinateur et trouver les mots-clés** qui correspondent bien au sujet à étudier.
- Ne pas hésiter à utiliser des **synonymes** ("**masse** éléphant Afrique" au lieu de "**poids**...", par exemple).
- Élargir le champ de recherches avec des termes **plus généraux** (pour obtenir plus de résultats).
- Affiner les résultats en ajoutant des termes **plus précis**.
- Ne pas faire des recherches contenant moins de **deux ou trois mots-clés** bien choisis simultanément.
- Certains mots trop commun qui pourraient figurer dans une requête sont écartés par les moteurs de recherche (**le, la, de, un, des**, ...). **Ils ne servent donc à rien.**

Les moteurs de recherche ne sont, à l'heure actuelle, **pas vraiment conçus pour reconnaître les langues naturelles**. Ni le français, ni l'anglais, ni même le chinois.
Il ne sert donc à rien de poser une question en langue naturelle.

➢ Varier les combinaisons de mots-clés utilisées

Ne pas hésiter à faire des recherches sur des variantes de combinaisons de mots-clés.

Pour un moteur de recherche, les deux expressions suivantes seront totalement différentes :

- hôtel côte belge
- vacances mer Nord

Pour un être humain (qui connaît la Belgique), il est clair qu'elles portent quasiment sur le même sujet.

➢ Attention aux accents

En principe, les caractères accentués sont sans importance. Une recherche sur "bébé phoque" donnera les mêmes résultats que la recherche sur "bebe phoque".

Attention toutefois si le mot existe dans une autre langue (c'est fréquent entre le français et l'anglais).

Faire une recherche sur le mot "elephant", par exemple, risque de ramener des résultats en anglais (ou les "e" ne sont pas accentués). Ce qui est ennuyeux si l'on ne maîtrise pas un peu cette langue.

➢ Les directives particulières

- **Singulier et pluriel**

On vérifie aisément que les recherches portant sur des mots au singulier ne donnent pas forcément les mêmes résultats que pour les mots au pluriel.

Exemple: faire une recherche sur "éléphant" et "éléphants". Conclure.

- **Synonymes**

Si l'on manque soi-même d'idées pour trouver des synonymes aux mots-clés de la recherche, on peut demander au moteur de recherche de faire le travail à notre place.

Exemple: faire une recherche sur "maison" et "~maison". Conclure. Fonctionne avec Google.

➢ Donner des directives aux moteurs de recherche

- **Exclure des mots**

Lorsqu'une recherche renvoie des résultats connexes qui ne sont pas souhaités, on peut demander à ce que certains mots soient évités:

- mettre un signe "-" devant le mot

Exemple: Je cherche des informations sur la biologie du Rossignol. Mais beaucoup de sites web évoquent le mot "Rossignol" sans aucun rapport avec les oiseaux. Je peux donc chercher

rossignol -ski -camping -village -"centre culturel" -chambres -domaine

- **Rendre un mot obligatoire**

Pour n'obtenir que les résultats qui contiennent obligatoirement un mot précis:

- mettre un signe "+" devant le mot

Exemple: **Rossignol +oiseau**
Exemple: **Python +reptile**

- **Rendre plusieurs mots obligatoires simultanément**

Pour obtenir les résultats qui contiennent en même temps plusieurs mots

- utiliser l'opérateur "**AND**" (qui signifie "ET", en français)

Attention, tous les moteurs de recherche n'acceptent pas cette directive. cfr BING

Exemple: **diamant AND émeraude**

- **Laisser le choix entre plusieurs mots**

Pour obtenir les résultats qui contiennent l'un ou l'autre des mots,

- utiliser l'opérateur "**OR**" (qui signifie "OU", en français)

Il faut obligatoirement utiliser l'opérateur en lettre majuscules.

Exemples:

- cheval OR chevaux
- Gand OR Gent (la ville de Belgique en français ou en néerlandais)

- **L'ordre des termes dans la requête peut être important**

Les premiers mots de la requête définissent le contexte, les mots suivants précisent le domaine de recherche.

Exemple:

- Pour chercher des informations sur un congrès de géographie se déroulant à Bruxelles, demander "géographie Bruxelles" plutôt que "Bruxelles géographie" (qui évoquera surtout la géographie de Bruxelles)

- **Faire une recherche dans un site web seulement**

Pour restreindre la recherche à un seul site web:

- ajouter l'expression **site:**www.lesite.com aux mots-clés de la recherche

Exemple: traitement texte site:sio2.be

- **Voir les sites liés à un autre**

Pour voir tous les sites qui ont établi un lien hypertexte vers un autre site:

- ajouter l'expression **link:**www.lesite.com aux mots-clés de la recherche

Exemple: **link:sio2.be**

- **Tableau des opérateurs numériques**

Types	*Fonction*	*Symboles*	*Exemples*
	Permettent des recherches selon des critères quantitatifs		
Opérateur égal à	**Recherche sur le nombre exact**	**=**	**=1998 : en 1998**
Opérateurs supérieur à, supérieur ou égal à…	**Recherche sur des périodes ou des séquences de nombres**	**>** **>=**	***> 1998* : depuis 1998** ***>=1998* : depuis 1998, y compris en 1998**
Opérateurs inférieur, inférieur ou égal à…	**Recherche sur des périodes ou des séquences de nombres**	**<** **<=**	**< 1998 : avant 1998** **<= 1998 : avant 1998 ou en 1998**
Opérateur d'intervalle	**Recherche entre deux dates, deux nombres…**	**:** **sur Google.com : nombre.. nombre**	**1995 :1998 : entre 1995 et 1998** **200€..300€ : sur Google.com, entre 200 et 300 €**

- **Tableau des opérateurs de proximité**

Types	*Fonction*	*Symboles*	*Exemples*
	Recherches sur le texte intégral selon la proximité des termes		
Opérateur d'adjacence	**Recherche sur des termes adjacents, dans l'ordre**		***Fibre ADJ optique* >> texte contenant**

	donné	**ADJ**	**l'expression " fibre optique "**
Opérateur de distance	**Recherche sur des termes séparés par une distance *n***	**nAV**	***Ecole 1AV privée* : >> école primaire privée, ou école technique privée**
Opérateur de proximité	**Recherche sur des termes présents dans le texte, quelle que soit leur distance**	**NEAR**	***Fibre NEAR optique* : >> texte contenant les deux termes, mêmes séparés**

- **Tableau des opérateurs de troncature**

Types	*Fonction*	*Symboles*	*Exemples*
	Substitution d'un symbole à des caractères, des lettres, des mots		
Troncature à droite	**Recherche sur tous les mots contenant la même racine ou le même préfixe**	*** + ?**	**franco* : >> francophone, francophonie, francophobe**
Troncature à gauche	**Recherche à partir d'un suffixe**	*****	***phobe : >> technophobe, agoraphobe, etc.**
Masque ou troncature centrale	**Remplace un ou plusieurs caractères dans un mot. Sur Google, remplace un mot dans une expression**	**? # sur Google : ***	**francopho?e > francophobe et francophone « le 21ème siècle sera * ou ne sera pas » > « le 21ème siècle sera *religieux, spirituel, laic…* ou ne sera pas »**

- **Tableau des autres opérateurs linguistiques**

	Fonction	*Symboles*	*Exemples*
Recherche	**trouver une expression**	**" ….. "**	***« document numérique »***

d'expression	**précise**		
Opérateur de définition	**Recherche de définitions dans différents sites**	**sur Google.com : define:**	***define:internet$*** **> donne des définitions d'internet**

7 Évaluation d'un site web

Les moteurs de recherche ne font aucune évaluation des sites web qu'ils indexent. Les "spiders" lisent tout et indexent tout. Sans faire de tri entre le bon et le moins bon.

Comment évaluer si un site web est fiable et si, pour moi, il correspond à ce que je peux attendre?

Pour répondre à ces questions, on peut utiliser la méthode "Qui - Quoi - Quand - Où - Pourquoi - Comment".

Pour évaluer la qualité d'un site web, nous allons nous poser ces 6 questions:

- **Qui?**

L'auteur du site web est-il identifiable? Est-ce une personne particulière ou s'agit-il d'un site d'une institution, d'un organe de presse, d'un gouvernement,...

Existe-t-il un moyen d'entrer en contact avec l'auteur, une adresse mail, un formulaire,... ?

S'il s'agit d'une personne particulière, quelles sont ses qualifications pour aborder le sujet? Est-ce un spécialiste dans le domaine? Est-il reconnu? Son nom est-il cité par d'autres sites web? Que dit-on de lui?

S'il s'agit d'un organisme ou d'une institution, son nom est-il connu? D'autres sites y font-ils référence?

Pour retrouver le nom et des informations sur l'auteur, il est parfois nécessaire de remonter jusqu'à la page d'accueil du site. Pour ce faire, il faut raccourcir progressivement l'adresse URL jusqu'au moment où l'on arrive à une page significative:

http://www.site.com/dossier1/sousDossierA/sousDossierX => le nom de l'auteur est-il accessible?

http://www.site.com/dossier1/sousDossierA/ => le nom de l'auteur est-il accessible?

http://www.site.com/dossier1 => le nom de l'auteur est-il accessible?

http://www.site.com/ => le nom de l'auteur est-il accessible?

Lorsque le site est produit par un organisme reconnu, le nom de cet organisme figure généralement dans l'adresse URL du site. Exemple: http://public.web.cern.ch/public/fr/lhc/Computing-fr.html

où l'on trouve la mention du CERN (Centre Européen pour la Recherche Nucléaire)

- **Quoi?**

Quelles sont les informations intéressantes que je trouve sur ce site?

Les informations données sont-elles compréhensibles? Sont-elles d'un niveau trop élevé? Sont-elles d'un niveau assez élevé?

Le document apporte-t-il des informations nouvelles? S'agit-il éventuellement d'une copie d'un autre site web (plagiat)?

Si les informations trouvées sont d'un niveau trop élevé par rapport à mes connaissances, je ne pourrai sans doute pas les exploiter. Si le site web est à l'usage des enfants de l'école primaire, son niveau ne sera peut-être pas suffisant.

Si les informations présentées me sont déjà connues, est-ce parce que je les ai déjà trouvées ailleurs? De nombreux sites se contentent de copier/coller des textes publiés ailleurs. Il faut alors retrouver l'original.

- **Quand?**

Quelle est la date de création du site?

Quelle est la date de dernière mise à jour de la page consultée?

Dans de nombreux cas, la date de publication des informations est importante. Peut-être les informations sont-elles obsolètes? L'auteur prend-il soin de remettre les informations à jour?

- ***Où?***

Pour quel endroit les informations données sont-elles pertinentes?

Si je cherche les tarifs postaux pour envoyer un colis, il faut m'assurer que les tarifs que je trouve sont valables dans mon pays. Si j'habite au Luxembourg, les tarifs en vigueur au Québec ne m'intéresseront pas.

- **Pourquoi?**

Pour quelle raison l'auteur a-t-il publié ce site web? A-t-il un intérêt financier, philosophique, religieux,... ?

Un site web d'une société qui produit des détergents de lessive pourrait avoir un intérêt à indiquer que ses produits sont inoffensifs pour l'environnement. Le site web d'une association de défense de l'environnement mettra plutôt en évidence des arguments dans la direction inverse.

Il faudra donc lire les informations données avec un esprit critique ou les écarter si l'on a des raisons de penser qu'elles peuvent ne pas être exactes et objectives.

- **Comment?**

La présentation du document est-elle correcte? Le texte est-il lisible? L'orthographe est-elle correcte? Les images présentées sont-elles intéressantes?

Il est possible de publier des informations très intéressantes et très pertinentes sans que la forme soit respectée. Mais, généralement, les sites web de qualité respectent le fond et la forme.

On accorde donc souvent une petite importance à la forme pour juger de la pertinence du fond.