

Introduction à internet

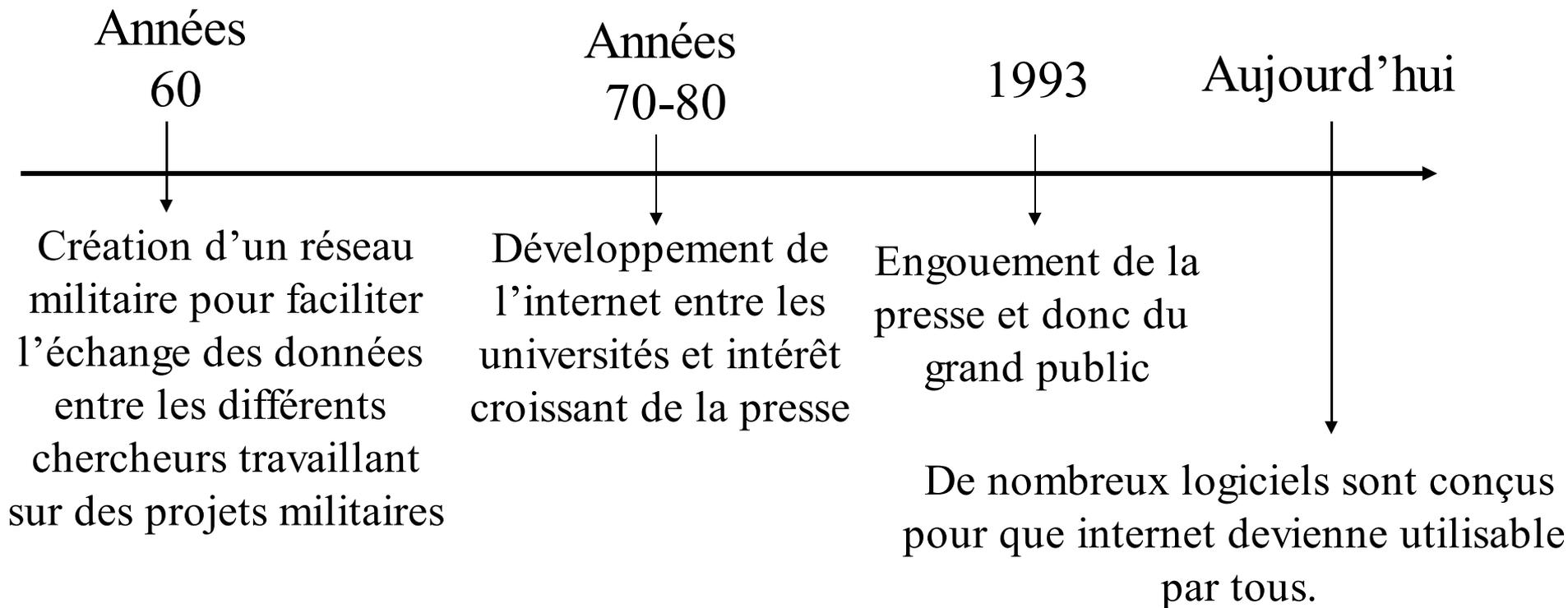
Chapitre 1 : Surfer sur le web

Surfer sur le web

- 1) Présentation de l'internet
- 2) Connexion à l'internet
- 3) Le monde du World Wide Web
- 4) Navigation sur le web
- 5) Web et sécurité

1) Présentation de l'internet

Un peu d'histoire ...



Qu'est-ce qu'internet ?

Systeme de plusieurs ordinateurs reliés en vue de partager des informations

Internet = un réseau mondial d'ordinateurs à accès public

50 à 80 millions d'internautes

Les informations véhiculées sur le web :
texte, images, son, vidéos, ...

Les services offerts : le courrier électronique,
discussions, WWW, telnet, ftp, news, ...

Ouvert à tous par le biais d'un abonnement dont le prix diminue sans cesse

2) Connexion à internet

L'équipement nécessaire

4 types de connexions :

- connexion permanente;
- connexion directe;
- connexion en mode terminal;
- connexion au courrier uniquement.

Equipement nécessaire

- Un ordinateur assez rapide pour que la connexion ne soit pas trop lente;
- Un modem pour convertir les informations lues ou envoyées sur la ligne téléphonique;
- Une ligne téléphonique;
- Un ou plusieurs logiciels pour gérer la connexion.

Aujourd'hui, de nouvelles techniques commencent à être implémentées pour accélérer le transfert des données :
câble (ADSL), lignes satellites, ...

Le modem n'est alors plus nécessaire.

La connexion permanente

- L'ordinateur est directement relié à un réseau TPC/IP (Transmission Control Protocol/ Internet Protocol) faisant partie de l'internet;
- Par exemple : ADSL ou le câble TV;
- Avantage : accès très rapide car exclusif;
- Adapté pour les gros consommateurs d'internet;
- Dans les structures importantes (universités), un équipement spécifique est utilisé pour la connexion via une ligne téléphonique à gros débit;

La connexion n'en reste pas moins une connexion permanente car exclusivement réservée à l'accès à internet

La connexion téléphonique directe

- L'ordinateur est relié à un réseau PPP ou SLIP (Point-to-point Protocol ou Serial Line Internet Protocol) via une ligne téléphonique permanente;
- Connexion la plus courante aujourd'hui car bon rapport qualité/prix;
- Adapté pour les petits consommateurs d'internet;
- Actuellement, il existe de nombreux abonnements gratuits (contre l'affichage de bannières publicitaires) mais reste le coût des communications téléphoniques;
- Une fois connecté, tout se passe comme pour une connexion permanente (mais le téléphone est occupé !);
- Pour accéder à ce service, vous devez contacter un fournisseur d'accès.

La connexion téléphonique en mode terminal

- L'ordinateur doit aussi contacter un fournisseur d'accès, mais tous les programmes utilisés fonctionnent chez le fournisseur d'accès;
- Avant 94, c'était le principal type de connexions, mais il est moins agréable car ne permet pas les interfaces graphiques => commandes indigestes;
- Aujourd'hui ce type de connexion a tendance à disparaître.

La connexion courrier

- Ce type de connexion permet uniquement d'envoyer et de recevoir du courrier électronique et dans certains cas d'accéder aux groupes de discussion;
- Ce n'est pas réellement une connexion à internet;
- Certaines entreprises l'utilisent encore en interne.

3) Le monde du World Wide Web

!! Attention !!

Web \neq Internet



Système de logiciels
qui s'exécutent sur
Internet

Réseau d'ordinateurs

Qu'est ce que le web ?

- **Principe de base = l'hypertexte** : en cliquant sur du texte mis en évidence, il est possible d'afficher un autre document;
- L'hypertexte existe depuis longtemps mais auparavant, les systèmes étaient limités : un clic donnait accès à un document contenu sur le même ordinateur, souvent sur le même disque et dans le même répertoire;
- Le WWW repousse ces limites : un clic peut déclencher l'affichage d'un document situé dans un autre pays, voire sur un autre continent sans l'autorisation de la personne à l'origine du document.

Comment accéder au web ?

- L'outil pour accéder au web est le navigateur web;
- 2 types de fonctionnement :
 - Les serveurs web qui s'exécutent sur des ordinateurs directement reliés au web:
 - Les navigateurs web qui s'exécutent sur des ordinateurs personnels et qui interrogent un serveur pour afficher les pages demandées.
- Les 2 principaux navigateurs sont Netscape Navigator (15% du marché) et Internet Explorer (85% du marché);
L'un n'est pas meilleur que l'autre mais leur compétition évite la situation de monopole.
- Il existe une dizaine d'autres navigateurs mais tous ne se valent pas; par sécurité, choisissez un des deux précédents.

Comment se procurer un navigateur ?

- Souvent les fournisseurs d'accès en proposent un;
- Les dernières versions des navigateurs sont disponibles sur le web : <http://www.netscape.com>
<http://www.microsoft.com>

La page d'accueil du navigateur

- C'est la page que le navigateur ouvre quand vous le lancez;
- Cette page peut-être personnalisée et doit idéalement contenir les liens que vous utilisez le plus souvent;
- Pour personnaliser la page d'accueil, il faut utiliser le langage de publication des pages web : HTML.

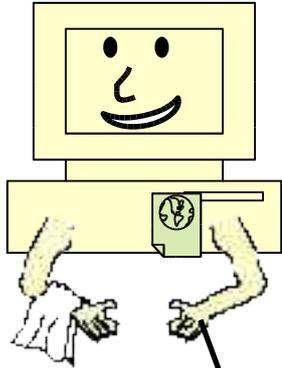
Un premier voyage sur le web

- Lorsque vous cliquez sur un lien, le navigateur envoie un message à un serveur web pour lui demander de rapatrier la page correspondante;
- Si le serveur est mis en service et n'est pas saturé, le document est transféré et affiché dans le navigateur;
C'est cette étape de transfert qui prend du temps;
- Pour pouvoir trouver des informations intéressantes, les navigateurs sont pourvus de liens renvoyant à des pages web essentielles sur la page d'accueil;
- L'adresse des liens est communiquée au navigateur dans le fichier source de la page web rédigé en mode texte HTML;
Il est d'ailleurs possible d'afficher le code source de toute page web;

Le serveur distant est surchargé.

Je vous l'envoie.

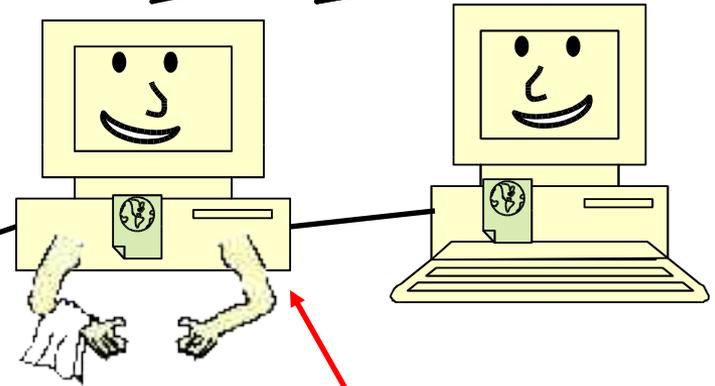
serveur FUNDP



Merci !

Qu'est-ce que je vous envoie ?

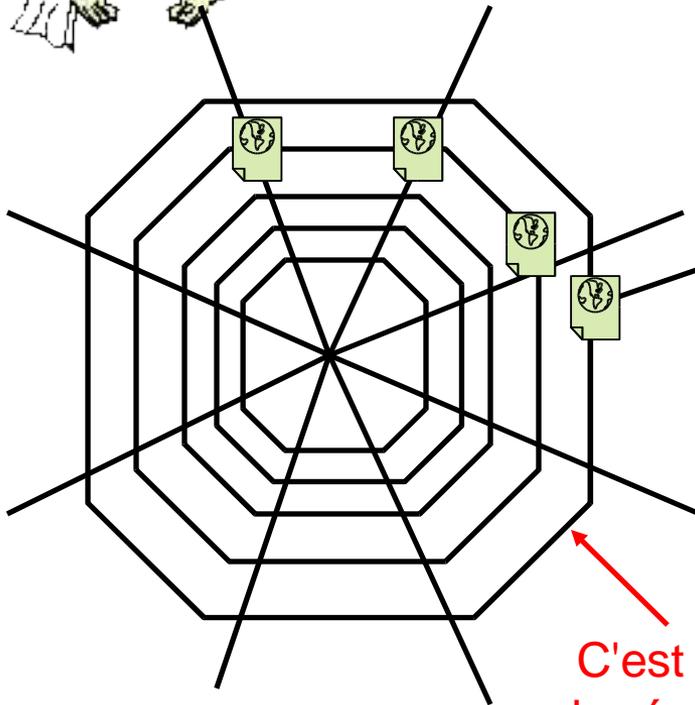
Maintenant je peux l'afficher !



Mon serveur est surchargé.

L'affichage est lent ?

C'est l'heure de pointe, le réseau est surchargé.



Comment s'y retrouver ?

Pour retrouver les pages déjà visitées :

- Précédent : affiche la page précédente;
- Suivant : affiche la page suivante si on a fait un retour en arrière;
- Accueil : affiche la page d'accueil;
- Signets ou Favoris : pour affecter un signet aux pages que vous consultez régulièrement;
- Historique : liste des pages déjà visualisées.

Liaison directe avec une URL

- URL=Uniform Ressource Locator
- Il est possible de taper directement l'adresse URL du site à visiter dans la barre de navigation;
- [http://](#) indique au navigateur qu'il s'agit d'une page web (Hyper Text Transfert Protocol)
- [www.nomsite.ext](#) est le nom d'hôte qui abrite le serveur web qui administre le site web;
- [/repertoire/](#) indique le nom du répertoire sur le serveur;
- [page.html](#) est le nom de la page.

<http://www.nomsite.ext/repertoire/page.html>

Autres protocoles internet

- [ftp://](#) pour le transfert de fichiers via un serveur à accès public;
- [telnet://](#) pour se connecter à un site telnet;
- [mailto:](#) pour envoyer un mail via le navigateur;
- [C:/Program/File/page.html](#) pour accéder à un fichier localement sur votre disque dur;
- ...

4) Naviguer sur le web

Ouvrir plusieurs fenêtres dans le navigateur

- Pour accéder à l'information plus rapidement ou pour ne pas perdre la trace d'une autre page web, on peut ouvrir plusieurs fenêtres (clic droit de la souris);
- **MAIS cela pose souvent problème** car chaque session utilise une portion de la mémoire et la vitesse de chargement globale est limitée;

Votre disque dur comme serveur web

- Lorsque le navigateur charge une page web, il la copie dans la mémoire cache. Lorsque le cache est plein, les données les plus anciennes sont effacées au profit des nouvelles;
- Chaque fois que vous demandez au navigateur de charger une nouvelle page, il examine son cache pour voir si la page ne s'y trouve pas;
- Les pages de la cache sont accessibles plus rapidement par exemple quand vous reculez dans l'historique
- Si au cours de votre session, vous passez souvent à la même page, elle finira par être sur votre disque dur, donc accessible encore plus rapidement;

Le cache en pratique

- Pour bénéficier des avantages de la cache, il est nécessaire de configurer le navigateur :
Affichage-Options-Général-Avancées-Paramètres
- Cache en mémoire = taille de la cache en mémoire centrale (RAM);
- Cache sur disque = taille de la cache sur le disque dur;
- Vérifier les documents = pour actualiser la version des documents sur le disque.

Le bouton Actualiser ou Recharger

- Lorsqu'une page est sauvée sur le disque, elle est figée et on ne peut plus voir les évolutions proposées par son concepteur;
- Le bouton Actualiser ou Recharger sert à actualiser le fichier en rechargeant la page web d'origine.

La commande Rechercher

- Parfois les pages web sont très longues, et il n'est pas évident de trouver l'information recherchée;
- La commande Rechercher permet de faire une recherche dans le document courant et non sur le site web;

Le clic droit de la souris (sur Mac, souris enfoncée longtemps)

Le clic droit donne directement la possibilité de

- Copier l'adresse d'un lien;
- Ouvrir le lien;
- Ouvrir le lien dans une nouvelle fenêtre;
- Enregistrer la cible ou le lien;
- Ajouter un signet ou favoris à la page en cours;

Capturer des fichiers sur le web

- Pour les stocker sur le disque dur : pour les lire plus tard ou parce qu'il s'agit de programme à exécuter;
- Pour les visualiser tout de suite, à condition de disposer du logiciel adapté pour les lire;
- Souvent les fichiers à télécharger sont compressés pour prendre moins de place, et doivent être décompressés avant d'être lus (ex. Winzip);

5) Sécurité sur le web

- Le principe du web est que toute information est accessible à tous; il ne faut donc pas diffuser n'importe quoi sur le web et surtout ne pas télécharger n'importe quel fichiers;
- Les données communiquées dans des formulaires ne doivent en aucun cas être des données confidentielles;
- Il existe cependant des méthodes pour crypter les données pour qu'elles ne puisse pas être lues par tous;
- L'adresse URL des serveur sécurisés commence par <https://> et le navigateur affiche un cadenas fermé dans le coin inférieur droit de la fenêtre;
- Les informations rapatriées sur le serveur subissent un codage avant leur émission de façon à être illisibles pour tout intermédiaire;

Introduction à internet

Chapitre 2 :

Le courrier électronique

Chapitre 2 : le courrier électronique

- 1) Configurer sa messagerie
- 2) Utiliser la messagerie
- 3) Bien utiliser la messagerie
- 4) Crypter les messages
- 5) Les groupes de discussion

1) Configurer sa messagerie

Les différents types de messageries:

- Netscape propose son propre système de messagerie;
- Les dernières versions de Windows proposent de travailler avec Outlook Express;
- Mais il est aussi possible d'utiliser un autre logiciel spécialement installé pour gérer le courrier électronique (par exemple Eudora <http://www.qualcomm.com>).

Le plus facile est de commencer avec la messagerie proposée par votre fournisseur d'accès.

Les serveurs de courrier POP et les autres

- Souvent les serveurs de courrier sont des serveurs POP (Post Office Protocol); ils mémorisent les messages qui leur sont envoyés jusqu'à ce qu'ils soient récupérés par leur destinataire;
- Il existe d'autres types de serveurs, mais il est nécessaire d'utiliser un serveur POP pour installer un logiciel de messagerie plus sophistiqué;
- Les serveurs de courrier IMAP (Internet Message Access Protocol), très répandus dans les entreprises qui possèdent une liaison directe avec internet, ne le permettent pas;
- Par exemple, Eudora permet de filtrer les messages automatiquement pour exclure les messages publicitaires et autres;

L'adresse électronique

`nom_de_compte@nom_de_domaine`

- Le nom de compte est le nom que vous utilisez pour vous connecter chez votre fournisseur d'accès;
- Le signe @ permet de séparer le nom de compte du nom de domaine;
- Le nom de domaine est en fait l'adresse de votre société, de votre fournisseur d'accès ou de votre service en ligne;
- Remarque :
nom de compte = user ID = nom d'utilisateur = pseudonyme

L'initialisation de la messagerie

Pas de panique ! Souvent, l'initialisation de la messagerie s'effectue de manière automatique par le logiciel de messagerie;

Vous pouvez avoir à définir les paramètres suivants :

- **Nom de la boîte à lettres POP** : c'est le nom de votre boîte aux lettres, communiqué au serveur pour qu'il vous transmette votre courrier; souvent c'est le nom que vous utilisez pour vous connecter chez votre fournisseur d'accès;
- **Serveur SMTP** (Simple Mail Transfer Protocol) : éventuel autre programme de messagerie utilisé pour envoyer votre courrier;
- **Mot de passe** : pour que vous seul puissiez accéder à votre courrier; parfois non nécessaire car prédéfini comme votre mot de passe utilisé pour la connexion internet;

- **Nom de l'utilisateur** : c'est votre nom complet qui est souvent joint de manière automatique à votre adresse;
- **Adresse de réponse** : si vous souhaitez que la réponse à votre courrier arrive à une autre adresse;
- **Options** :
 - fixer la fréquence d'interrogation de la boîte aux lettres;
 - choisir la police d'affichage, sa taille, ...
 - inclure automatiquement le message original dans la réponse;
 - inclure automatiquement une signature à la fin de tous vos messages;
 - ...

2) Utiliser la messagerie

L'envoi d'un message

- Si vous ne savez à qui envoyer votre premier message, vous pouvez, l'envoyer à votre propre adresse !
- Après avoir lancé votre programme de messagerie, choisissez « Nouveau message » ou « Rédiger un courrier »;
- Après avoir rédigé votre message, choisissez « Envoyer » ou « Expédier »; s'il y a une erreur dans l'adresse du destinataire, un message vous reviendra du serveur, disant qu'il n'a pas pu le délivrer;
- Remarquez qu'il n'est pas nécessaire d'être connecté pour rédiger le message; souvent, le bouton « Envoyer » permet la connexion directe le temps de l'envoi;

- Quelques éléments que vous pourrez trouver dans la fenêtre de composition d'un courrier :
 - **A (To)** : c'est l'adresse de la personne à qui vous voulez écrire;
 - **De (From)** : s'il n'est pas fourni de façon automatique, ce champ vous permet de préciser votre adresse à votre correspondant;
 - **Réponse à** : permet de différencier l'adresse d'où provient le message (De) de l'adresse de réponse;
 - **Objet (Subject)** : c'est le titre du message, qui apparaîtra dans la boîte aux lettres;
 - **Cc (Copie Carbone)** : l'adresse d'un autre correspondant à qui le message est aussi destiné;
 - **Cci (Copie Carbone Invisible)** : pour que le destinataire principal ne sache pas que le message est aussi envoyé à cette autre adresse;
 - **Pièces jointes (Attachments)** : pour joindre un fichier au message (détaillé plus loin);
 - **Le corps du message;**

La réception d'un message

- Souvent, vous êtes informé qu'il y a des messages dans votre boîte aux lettres dès votre connexion;
- Vous obtenez les différents messages en sélectionnant une commande du type « Obtenir les nouveaux messages » ou « Recevoir »;
- Vous pouvez choisir de « Répondre » directement à votre correspondant et son adresse est alors automatiquement complétée dans la fenêtre d'édition de message;
- Dans le message de réponse, il vous est possible d'« Inclure » le message original, qui sera alors mis en évidence par exemple précédé de >;

L'envoi de fichiers

Les fichiers qui transitent sur le net utilisent un des trois codages suivants :

3. UUENCODE :

- quelque soit le type de fichier, il est converti en texte pur par un processus d'uuencodage;
- si la messagerie du destinataire convertit automatiquement les fichiers uuencodés, le fichier sera lu facilement;
- sinon le message doit être sauvegardé et uudécodé pour être lu.
- ce système a tendance à disparaître au profit du codage MIME;

1. BinHex :

- processus très proche du UUENCODE mais utilisé par les ordinateurs Macintosh;
- ce processus a tendance à disparaître car de plus en plus de messagerie Mac utilisent les formats MIME et UENCODE;

2. MIME (Multimédia Internet Mail Extension) :

- processus de codage qui facilite le transfert de fichier;
- la messagerie identifie le type du fichier et le décode directement de sorte que les utilisateurs ne s'aperçoivent plus du codage;
- les logiciels de messagerie les plus populaires utilisent ce type de codage;

En pratique ...

- sachez que les fichiers de texte sans mise en page peuvent être lu par tous (.txt,.rtf,...);
- les fichiers Word, Excel, ... peuvent poser problème car ils contiennent du texte avec mise en page mais il est possible de sauvegarder ces documents au format texte;
- le type d'ordinateur et le type de messagerie sont déterminants de la qualité du transfert;
- choisissez une messagerie récente et recommandez à vos correspondants de faire de même, la tendance est de travailler avec le codage MIME qui est plus universel;
- si les fichiers ne passent pas, il est possible de télécharger des logiciels de conversion via le web;
- si vous envoyez des images par mail, préférez les formats .gif et .jpeg qui sont moins lourds que .bmp,.tif, ...

3) Bien utiliser la messagerie

Quelques suggestions pour mieux utiliser la messagerie :

- **Carnet d'adresses** : disponible dans toute les messageries, il permet de mémoriser les adresses et de les retrouver à partir du nom réel de la personne;
- **Alias** : raccourcis pour une ou plusieurs adresses;
- **Publipostage** : création de listes d'adresses électroniques pour envoyer un même courrier à toutes les adresses de la liste;
- **Travailler déconnecté** : pour écrire et lire les messages, il n'est pas nécessaire d'être connecté, il suffit de se connecter le temps d'envoyer et de réceptionner les messages;

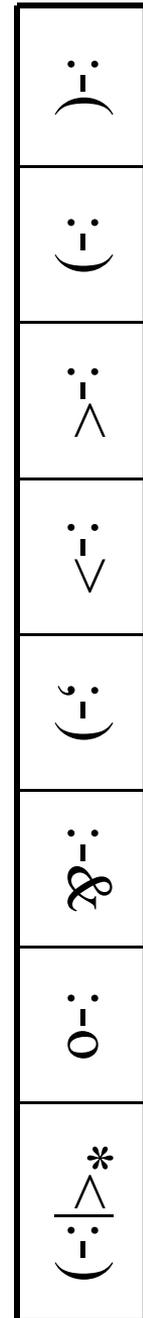
- **Faire suivre ses messages** : si vous travaillez à la maison et au bureau avec des messageries différentes, il est possible que tous vos messages du bureau soient automatiquement envoyés vers votre adresse à la maison;
- **Signaler son absence** : en cas d'absence, il est parfois possible d'envoyer un message automatique à toutes les personnes qui essaient de vous contacter;
- **Filtrer les messages** : les messageries sophistiquées permettent de filtrer les publicités et de réorganiser les messages selon leur origine;

Toutes les messageries ne permettent pas toutes ces fonctionnalités, mais il est recommandé d'utiliser celles qui sont proposées.

Quelques touches d'originalité

- les smileys

:-(:-) :-< :-> ;-) :-& :-o *< :-)	pas content content triste grand sourire clin d'oeil je ne sais pas quoi dire je suis sous le choc le père Noël
---	--



- les abréviations

ASAP	As soon as possible
DQP	Dès que possible
FYI	For your information
IMHO	In my humble opinion
AMHA	A mon humble avis
BTW	By the way
TIA	Thanks in advance
ROTFL	Rolling on the floor (C'est à se tordre de rire)

Les dangers de la messagerie

- Vos messages peuvent être détournés par un pirate informatique, à moins qu'ils ne soient cryptés;
- Votre destinataire peut faire suivre votre message à une autre personne sans vous en demander la permission;
- Attention aux messages écrits sous l'emprise de la colère, ils risquent de vous causer des ennuis et ils sont souvent mal compris par le destinataire;
- Du point de vue juridique, les messages électroniques n'ont pas de valeur car ils sont falsifiables; mais le témoignage de plusieurs personnes concernant un message reçu peut être dangereux;

Les virus informatiques

virus = programme qui se loge dans un ordinateur pour faire des dégats; souvent, il s'auto-propage par mail;

- seuls les fichiers capable d'effectuer des actions peuvent contenir des virus : programmes (.exe) ou document texte écrit avec un traitement de texte qui possède un langage macros (Word et Excell par exemple);
- impossible donc de contracter un virus par un fichier texte ou une image;
- par précaution, évitez d'ouvrir les messages de destinataires inconnus, même si le titre vous paraît anodin;
- il est donc important de se munir d'un bon anti-virus et de le mettre à jour régulièrement (c'est aujourd'hui possible via le web);

Les hoax ou canulars

hoax = message incité à être propagé à grande échelle pour une bonne cause en fait inventée;

- Ces messages encombrant les liaisons car ils peuvent faire plusieurs fois le tour du monde, mais ils sont inoffensifs;
- Particularités : description d'une situation urgente, utilisation de termes techniques ou du nom d'une association connue pour valider la démarche, pas de date d'émission;
- Le meilleur site francophone de détection des hoax : <http://www.hoaxbuster.com>, à consulter avant de faire suivre n'importe quoi;

Et si vous receviez un message du type ...

- PUB** • « Commandez Viagra online »
- VIRUS** • « Pamela Anderson nue » + un fichier .zip
- HOAX** • « Urgent : don de sang »
- VIRUS** • « Le fichier que vous m'avez demandé » + un fichier Excell
- HOAX** • « Aidez-moi à retrouver ma fille » + un fichier .gif
- PUB** • « Commandez notre nouveau logiciel »
- VIRUS** • « Offre gratuite de la nouvelle version » + un fichier .exe
- HOAX** • « Totem Mantra », propose une chaîne de l'amitié
- HOAX** • « Microsoft a détecté un nouveau virus »
- VIRUS** • « Blague : blanche-neige et les 7 nains » + un fichier .doc

... effacez -le

4) Le cryptage du courrier

Principe du cryptage

- Avant d'être envoyé, le message est crypté selon un procédé connu par vous et votre destinataire;
- Pendant son transfert, personne ne peut plus lire le message, savoir d'où il vient, ni où il va;
- Il existe 2 types de codage :
 - le codage par clé privée;
 - le codage par clé publique.

Le codage par clé privée

- Principe : rendre le fichier informatique inutilisable sans un mot de passe;
- Le nom du fichier et le mot de passe sont communiqués au programme de cryptage pour rendre le fichier inutilisable sans le mot de passe;
- C'est un moyen simple de crypter un fichier, mais il existe des algorithmes rapides pour casser les clés publiques;
- Si vous utilisez ce type de clé, et de façon générale quand vous choisissez un mot de passe, évitez les mots du dictionnaire et composez un mot avec des lettres, des chiffres et des symboles spéciaux;

Le codage par clé publique

- Principe : le document est crypté à l'aide d'une clé publique et doit être décrypté avec une clé privée;
- La première clé est publique et peut donc être communiquée à tous; il existe d'ailleurs des répertoires des clés publiques et elles peuvent être attachées au courrier;
- Le message est codé à l'aide de la clé publique du destinataire et son décryptage nécessite la connaissance de la clé privée du destinataire;
- La taille de la clé détermine le niveau de sécurité : le décryptage d'une clé 40 bits suppose des moyens colossaux et le décryptage d'une clé 128 bits prendrait un temps quasi infini;
- En pratique, tout se fait automatiquement, du moins dans les messageries récentes;

5) Les groupes de discussion

Les groupes de discussion

- groupes de personnes s'échangeant des messages sur un sujet bien précis;
- il existe des listes de discussion sur tous les sujets : les loisirs, les logiciels informatiques, les fans d'un groupe, ...
- attention, on peut passer beaucoup de temps dans les groupes de discussion !
- il existe les groupes de discussions modérés, c'est-à-dire qu'une personne lit les messages avant qu'ils soient publiés;
- les principaux navigateur proposent l'accès à différents groupes de discussion sinon vous pouvez trouver des groupes publiques sur le web qui son alors appelés News;

Les listes de diffusion

- les listes de diffusion proposent aussi d'échanger des idées sur un sujet commun, mais tout message sera envoyé à tous les membres de la liste;
- il faut s'inscrire pour pouvoir participer à une liste de diffusion;
- souvent, les listes de diffusion sont automatisées sauf si elles sont modérées;
- attention, en vous inscrivant à une liste de diffusion, vous risquez de recevoir un nombre considérable de mail par jour !
- il est parfois pratique de s'inscrire un jour sur une liste de diffusion pour trouver la réponse à un problème précis;

Les forums ou les news

ce sont des groupes de discussion diffusés par l'intermédiaire d'un site web; les messages sont affichés dans le navigateur et pas envoyé vers les boîtes mail; ils sont donc publics;

Introduction à internet

Chapitre 3 : FTP et telnet

Chapitre 3 : FTP et Telnet

- 1) FTP
- 2) Telnet

Référence intéressante sur ftp :

<http://help.yahoo.com/help/cf/geo/gftp/>

1) FTP

FTP = File transfert Protocol

- sur l'internet, de nombreux fichiers peuvent être téléchargés gratuitement;
- le moyen le plus sûr de téléporter des fichiers est d'utiliser ftp, qui a été conçu à l'origine pour le transfert de fichier entre machines UNIX, avant que le courrier électronique n'existe;
- certains sites ftp sont protégés par un mot de passe mais il est souvent possible de se connecter comme « anonymous » et d'accéder à une partie des données;
- aujourd'hui, il est possible de se connecter aux serveurs ftp via le navigateur web (ftp:// ...);
- sur les Mac, ftp s'appelle fetch !

Les types de fichiers

- presque tous les types de fichiers peuvent être transférés par ftp;
- évitez les majuscules et les accents dans les noms de fichiers;
- les fichiers textes et html peuvent être transférés avec les options « ASCII » et « binary »;
- les autres fichiers doivent être transférés avec l'option « binary » ou « raw data »;
- souvent, une option « auto » est proposée qui détecte automatiquement le type du fichier et le type de transfert adapté;

Les fichiers compressés

- la compression de fichier permet de gagner 40 à 75% de taille sauf éventuellement pour les exécutables;
- les types de fichiers compressés :

.zip	sous Dos, Windows et Mac
.exe	archive auto-extractible sous Dos ou Windows
.sea	archive auto-extractible sous Mac
.gz .shar .tar .Z	Sous UNIX
.hqx .pit .sit	sous Mac

L'accès aux fichiers

- dans le navigateur, les répertoires et noms de fichiers apparaissent en temps que liens hypertexte;
- pour localiser le fichier recherché, ce n'est pas toujours simple, mais les fichiers INDEX, README peuvent vous aider;
- les commandes « get » et « put » permettent de rapatrier ou d'envoyer le fichier au serveur;
- parfois, ces commandes sont symbolisées par des flèches de gauche à droite ou de droite à gauche (WS_FTP);

La recherche sur les sites ftp

- Il existe des programmes de recherche qui proposent de trouver des fichiers ftp sur un ou plusieurs serveurs spécifiés (par exemple <http://alltheweb.com>);
- la recherche d'information sur le web est détaillée dans le chapitre suivant;

2) Telnet

- Telnet = programme qui permet d'accéder à une machine distante protégée ou non par un mot de passe;
- Utilisation : lors d'un séjour à l'étranger ou simplement pour consulter des données ou tester un programme;
- A l'aide du navigateur web, il est possible de se connecter à un serveur telnet;
- La suite des opérations se déroule alors sur le serveur distant et plus sur votre ordinateur;
- Le temps de réponse est parfois très long, selon l'encombrement des lignes;

Introduction à internet

Chapitre 4 :

Rechercher de l'information sur le web

Chapitre 4 : Rechercher de l'information sur le web

- 1) Le web invisible
- 2) La recherche intuitive
- 3) Les outils de recherche
- 4) Les étapes d'une recherche efficace
- 5) Les spécificités de certains moteurs de recherche
- 6) Les messages d'erreur
- 7) Les aspirateurs de sites

1) Le web invisible

- Les meilleurs moteurs de recherche indexent moins de 20% du web;
- En effet, les outils de recherche ne référencent pas
 - les pages non html;
 - les pages non référencées;
 - les fichiers dynamiques;
 - les pages protégées par un mot de passe;
 - les bases de données;

2) La recherche intuitive

Lorsqu'on recherche un site dont on connaît l'institution, la société ou l'organisation, il est relativement facile de deviner l'adresse de ce site :

- `http://www.`
- le nom de ou l'acronyme de l'organisation
- le domaine le plus évident :
 - `.com` pour une société commerciale;
 - `.org` pour une organisation internationale;
 - `.be` pour un site belge;
 - `.fr` pour un site français;

La recherche intuitive permet aussi de retrouver une page qui n'existe plus ou des liens morts en coupant l'URL jusqu'à chaque `/`.

3) Les outils de recherche

- Les moteurs : pour une recherche précise
- Les annuaires : pour une recherche large
- Les métamoteurs : pour dégrossir rapidement
- Les encyclopédies : pour des infos permanentes et sûres
- Les sites fédérateurs et les guides : pour une recherche large dans un domaine précis
- Les webrings : pour une recherche large dans un domaine précis

Les moteurs : pour une recherche précise

- Les moteurs de recherche indexent régulièrement des millions de pages web dans une base de données à l'aide de robots qui inventorient et ajoutent les nouvelles pages;
- Chaque robot fonctionne à sa manière et ils peuvent donc produire des résultats très différents selon l'indexation utilisée :
 - par mots clés (balise META keywords dans le document HTML)
 - par titre (balise TITLE dans le document HTML)
 - dans le texte (tous les mots ou les mots des premières lignes)
- Chaque moteur propose un formulaire de recherche avancée plus puissant;

- Le moteur recherche l'occurrence du ou des mots dans sa base de données (souvent 20% du web);
- Les résultats sont classés selon un algorithme de pertinence propre (présence du mot dans le titre, dans le texte, dans les balises métas, indice de popularité, ...)
- Les résultats sont souvent présentés après clustering, c'est-à-dire qu'ils ne proposent qu'un seul lien pour plusieurs pages sur le même site;
- Exemples : google, alta vista, hot bot, ...

Les annuaires : pour une recherche large

- Les annuaires recensent des sites classés par catégories et sous-catégories; le classement est effectué par la main de l'homme et la pertinence et la qualité des sites sont donc bien meilleures;
- Les responsables construisent une base de données de fiches descriptives des différents sites et la recherche des mots s'effectue sur base de la fiche descriptive et non du site;
- Outre la recherche par catégorie, il est possible de réaliser une recherche par mots clés qui fournira les sites et les différentes catégories pertinentes;
- Exemples : yahoo, ...

Les métamoteurs : pour dégrossir rapidement

- Ils permettent d'interroger simultanément plusieurs moteurs de recherche mais la requête doit donc être simple;
- Exemples : ariane6, metacrawler, ...

Les encyclopédies : pour des infos permanentes et sûres

- Elles présentent une source sûre, les informations sont de qualité et permanentes; elles sont parfois payantes;
- Exemples : britannica, quid, xrefer, ...

Les sites fédérateurs et les guides : pour une recherche large dans un domaine précis

- Les sites fédérateurs et passerelles thématiques sélectionnent des sources de qualité dans un domaine précis;
- Il sont gérés par des professionnels ou par des passionnés de l'information et ouvrent des liens vers le web invisible;
- Exemple : les universités en proposent souvent;

Les webrings : pour une recherche large dans un domaine précis

- Il s'agit d'une sélection de sites traitant d'un thème donné et reliés entre eux;
- Exemple : dans tous les domaines ...

Veillez à avoir quelques favoris avec chacun de ces outils de recherche, vous pourrez ainsi choisir le meilleur outil pour faire une recherche d'information sur le web.

=> remplissez bien votre boîte à outils

4) Pour une recherche efficace ...

Préparer sa recherche

- Délimiter le sujet et choisir les mots clés;
- Choisir le type de document attendu;
- Choisir les limites de ma recherche (langue, date de publication, zone géographique des pages, ...);
- Fixer le temps de recherche et la quantité d'information attendue;
- Définir au mieux la question que l'on se pose;
- Choisir l'outil le mieux approprié pour la recherche;

Exprimer ses critères de recherche

- chaque outil de recherche possède un langage d'interrogation spécifique à maîtriser pour une recherche efficace;
- pour la plupart des outils, l'ordre des mots est sans importance, mais certains outils accordent un poids au mot en fonction de sa position;
- pour forcer la prise en compte de l'ordre des mots, il faut les placer entre guillemets;
- les moteurs de recherche excluent les mots fréquents (et, mais, si, ...), ce qui peut être problématique pour la recherche;
- la plupart des moteurs sont plus ou moins insensibles à la casse, mais sensibles aux accents (moteurs francophones);
- certains moteurs cherchent automatiquement tous les mots dérivés, pluriels, ...

- les **opérateurs booléens** permettent d'associer deux ou plusieurs mots pour une recherche plus efficace : **AND, OR et NOT**;
- sur certains moteurs, il existe un **opérateur de proximité (NEAR)** accessible à partir du formulaire de recherche avancée; il permet de rechercher deux mots proches l'un de l'autre;
- pour rechercher une **expression**, il suffit de la mettre **entre guillemets**;
- le **caractère générique (* ou ? ou %)** permet de remplacer n'importe quelle lettre ou chaîne de caractère;
- la **recherche par champs** est une fonction très utile et assez méconnue qui permet de restreindre la recherche à certaines parties des pages web (titre, URL, lien, ...);
- la **recherche imbriquée** permet d'utiliser plusieurs opérateurs à l'aide de parenthèses;
- le **langage naturel** consiste à poser une question sous forme de phrase au moteur; il interroge alors une base de données composée de questions dont il possède la réponse; cela peut-être très efficace;

Evaluer les ressources

- Vous savez maintenant comment trouver de l'information sur le web, mais il faut savoir que tous les sites ne proposent pas une information fiable;
- Quelques indices de crédibilité :
 - l'URL : elle donne le domaine d'activité;
 - la source du document : il doit être possible d'obtenir le nom de l'auteur, le nom de l'institution,...
 - l'actualité : la date de création et la date de la dernière mise à jour; à relativiser en fonction du sujet;
 - les références bibliographiques : si le contenu de la page n'est pas original, les sources originales doivent être indiquées;

Citer les références web

- Une page ou un site web utilisé pour rédiger un document doit être cité;
- Pour un site web, il faut donner l'auteur, le titre de la page d'accueil, l'adresse URL et la date de la consultation;
- Pour une page web, il faut donner l'auteur, le titre de la page, la ressource plus large à laquelle le document est rattaché, l'adresse URL et la date de consultation;
- Pour un document sur un site ftp, il faut donner l'auteur, le type de support, la date, l'adresse ftp, le répertoire et le fichier;

5) Les spécificités de certains outils

<http://www.searchengineshowdown.com>

Les moteurs de recherche :

- Alltheweb
- Alta Vista
- Google
- Hot Bot
- Lycos
- MSN
- Teoma
- WiseNut

Les petits nouveaux :

- Gigablast
- Openfind

Les annuaires :

- Britannica
- Looksmart
- Yahoo

Alltheweb

- Types de fichiers : html, pdf, doc, flash, news, images, vidéos, MP3, ftp;
- Opération par défaut : AND;
- Recherche booléenne : + pour AND, - pour NOT, () pour OR;
- Recherche de proximité : termes entre guillemets;
- Troncature : pas de troncature pour l'instant;
- Sensibilité à la casse : insensible;
- Recherche par champs : titre, URL, lien, ...
- Limites : langue (46), actualité (6 choix), type de fichier, domaine;

- Mots fréquents : ils sont recherchés;
- Tri : par pertinence (# de liens pointant vers cette page);
- Avantages :
 - large, rapide et adapté à une recherche précise;
 - sensible aux mots fréquents;
 - indexe la page web entière, les pdf, les doc et les fichiers flash;
 - la base de donnée est mise à jour régulièrement;
 - personnalisation des paramètres de recherche par défaut;
- Inconvénients :
 - pas d'archive en mémoire;
 - l'éventail de recherche est encore limité;

Alta Vista

- Types de fichiers : html, pdf, images, MP3, video, news;
- Opération par défaut : AND pour une recherche simple, OR pour une recherche avancée, mais cela varie;
- Recherche booléenne : + pour AND, - pour NOT, | pour OR, ! pour AND NOT, (et);
- Recherche de proximité : termes entre guillemets ou ~ pour NEAR;
- Troncature : avec le caractère générique * pour remplacer n'importe quoi;
- Sensibilité à la casse : insensible pour la recherche simple, sensible pour la recherche avancée;
- Recherche par champs : titre, URL, lien, ...;

- Limites : langue (beaucoup) , actualité (entre 2 dates), type de fichier;
- Mots fréquents : pris en compte pour la recherche simple, ignorés pour la recherche avancée;
- Tri : pertinence et clustering;
- Avantages :
 - beaucoup de possibilités dans la recherche avancée;
 - recherche de proximité, troncature, recherche de liens;
 - couverture internationale (beaucoup de langues);
- Inconvénients :
 - indexe les 110 premiers K d'une page web et 750 premiers K des pdf;
 - pas d'archives indexées.

Google

- Types de fichiers : html, pdf, ps, doc, xls, txt, ppt, rtf, asp, wpd;
- Opération par défaut : AND;
- Recherche booléenne : rien pour AND, - pour NOT, OR pour OR ;
- Recherche de proximité : termes entre guillemets;
- Troncature : pas de troncature, mais * remplace n'importe quel mot et ~ permet un recherche de synonyme;
- Sensibilité à la casse : insensible;
- Recherche par champs : titre, URL, liens, ...
- Limites : langue, domaine, actualité, type de fichier et contenu pour adulte;

- Mots fréquents : ils sont ignorés sauf si précédés d'un +;
- Tri : par pertinence avec une pondération selon le type de lien et par clustering ;
- Avantages :
 - **le plus vaste**, par la taille et le type de fichiers pris en compte;
 - le tri par pertinence avec la pondération donne des résultats excellents;
 - il prend en compte les archives en mémoire;
 - il permet la recherche des mots fréquents;
- Inconvénients :
 - l'éventail de recherche est limité : pas de troncature, ...
 - la recherche de liens est une recherche exacte;
 - il est difficile d'obtenir le tri sans clustering;
 - il indexe seulement les 101 premiers K d'une page web et les 120 premiers K d'un pdf;

Hotbot

(Open Directory + Direct Hit + Inktomi +Fast)

- Types de fichiers : html, image, video, MP3, Java, pdf;
- Opération par défaut : AND;
- Recherche booléenne : & ou + pour AND, | pour OR, ! ou - pour NOT, (et);
- Recherche de proximité : terme entre guillemets ou recherche de phrase;
- Troncature : le caractère générique * existe mais pose parfois problème s'il est utilisé ailleurs qu'en fin de mot;
- Sensibilité à la casse : oui;
- Recherche par champs : par la recherche avancée;
- Limites : actualité (entre 2 dates), langue (9), type de fichier, domaine, continent, profondeur du répertoire de résultat;

- Mots fréquents : vaste liste de mots fréquents qui sont ignorés;
- Tri : par pertinence et par clustering (facultatif);
- Avantages :
 - vaste éventail de recherche possible;
 - aide à la recherche avancée;
 - recherche rapide dans 4 grandes bases de données;
- Inconvénients :
 - ne permet pas d'utiliser toutes les possibilités de recherche des 4 bases de données;
 - il manque une recherche selon le type de fichier;

Lycos

- Types de fichiers : html, images, vidéo, ftp, MP3;
- Opération par défaut : AND;
- Recherche booléenne : + pour AND, - pour NOT;
- Recherche de proximité : terme entre guillemets ou option rechercher la phrase;
- Troncature : pas de troncature;
- Sensibilité à la casse : insensible;
- Recherche par champs : titre, URL, lien, ...
- Limites : limite de langue;

- Mots fréquents : ils sont recherchés;
- Tri : par pertinence;
- Avantages :
 - contenu élargi depuis peu;
- Inconvénients :
 - recherche booléenne partielle;
 - mise à jour lente;

MSN

(Looksmart + Inktomi + Direct It)

- Types de fichiers : html, pdf, audio, vidéo, images;
- Opération par défaut : AND;
- Recherche booléenne : AND pour AND, OR pour OR, NOT ou – pour NOT, (et);
- Recherche de proximité : termes entre guillemets ou option recherche de phrase;
- Troncature : pas de troncature mais l'option « word stemming » permet de rechercher les mots dérivés;
- Sensibilité à la casse : la casse usuelle est ignorée mais les majuscules en milieu de mots sont prises en compte;
- Recherche par champs : titre et lien;

- Limites : langue (14), type de document, profondeur du répertoire de résultats;
- Mots fréquents : beaucoup de mots fréquents qui sont tous ignorés;
- Tri : par pertinence + options : un résultat par domaine;
- Avantages :
 - une vaste base de données;
 - limite de profondeur de page;
 - recherche avancée très efficace;
- Inconvénients :
 - la recherche de lien doit être exacte;
 - la troncature pose souvent des problèmes;
 - pas d'archives;

Teoma

- Types de fichiers : html;
- Opération par défaut : AND;
- Recherche booléenne : - pour NOT, **OR pour OR**;
- Recherche de proximité : termes entre guillemets ou option recherche d'une phrase;
- Troncature : pas disponible pour l'instant;
- Sensibilité à la casse : la casse est ignorée;
- Recherche par champs : **titre, url, lien**;
- Limites : **langue (10) et domaine**;
- Mots fréquents : ignorés par défaut mais pris en compte si précédés de +;

- Tri : par pertinence mais chaque recherche donne lieu à 3 catégories de résultats (résultats, raffinements possibles, ressources);
- Avantages :
 - nouvelle base de données originale avec de nombreux sites sponsorisés;
 - présentation des résultats originale;
- Inconvénients :
 - mise à jour non régulière;
 - maigre éventail de recherche;
 - impossible de supprimer le clustering;

WiseNut

- Types de fichiers : html;
- Opération par défaut : AND;
- Recherche booléenne : - pour NOT;
- Recherche de proximité : termes entre guillemets ou option recherche d'une phrase exacte;
- Troncature : non disponible;
- Sensibilité à la casse : insensible;
- Recherche par champs : non disponible;
- Limites : langue (25);
- Mots fréquents : ils sont ignorés sauf si précédés de +;

- Tri : par pertinance ;
- Avantages :
 - nouvelle base de données;
 - personnalisation possible des paramètres de recherche;
- Inconvénients :
 - pas d'archives indexées;
 - manque de possibilités de recherches;

Gigablast

- Types de fichiers : html, pdf, doc, ppt, ps, xls;
- Opération par défaut : OR;
- Recherche booléenne : + pour AND, - pour NOT;
- Recherche de proximité : termes entre guillemets et possibilité de rechercher la phrase exacte;
- Troncature : non disponible;
- Sensibilité à la casse : insensible;
- Recherche par champs : titre, type, url, ...;
- Limites : type de fichier et domaine;

- Mots fréquents : ils sont recherchés;
- Tri : par pertinence ou date;
- Avantages :
 - tri par date
 - archives indexées;
- Inconvénients :
 - base de données réduite;
 - problèmes avec le OR;
 - possibilités de recherche limitées;

Openfind

- Types de fichiers : html;
- Opération par défaut : AND;
- Recherche booléenne : non disponible;
- Recherche de proximité : termes entre guillemets et possibilité de rechercher la phrase exacte;
- Troncature : non disponible;
- Sensibilité à la casse : insensible;
- Recherche par champs : non disponible;
- Limites : langue (chinois traditionnel et simplifié, japonais et coréen);

- Mots fréquents : ils sont recherchés;
- Tri : par pertinence, date ou taille;
- Avantages :
 - nouvelle base de données;
 - recherche en langue chinoise;
- Inconvénients :
 - nombreux résultats dupliqués et liens inactifs;
 - pas d'archives indexées;
 - possibilités de recherche limitées;

Britannica

ressources scolaires et académiques

- Types de fichiers : html;
- Opération par défaut : OR;
- Recherche booléenne : + ou AND pour AND, OR pour OR, NOT ou – pour NOT;
- Recherche de proximité : termes entre guillemets ou ADJ;
- Troncature : caractère générique * pour tronquer un mot et obtenir les mots dérivés;
- Sensibilité à la casse : sensible;
- Recherche par champs : seulement par titre;
- Limites : niveau de classement Britannica, domaine et continent;

- Mots fréquents : ils ne sont pas recherchés;
- Tri : par pertinence;
- Avantages :
 - orienté scolaire et académique;
 - recherche de certains périodiques;
- Inconvénient :
 - le répertoire n'est pas très vaste;

Looksmart

- Types de fichiers : html;
- Opération par défaut : AND;
- Recherche booléenne : non disponible;
- Recherche de proximité : non disponible;
- Troncature : automatique;
- Sensibilité à la casse : insensible;
- Recherche par champs : non disponible;
- Limites : non disponible;
- Mots fréquents : ils ne sont pas recherchés;
- Tri : par pertinence;
- Avantage : un des plus larges répertoires;
- Inconvénient : pas de recherche avancée;

Open Directory

- Types de fichiers : html;
- Opération par défaut : AND;
- Recherche booléenne : AND, OR, ANDNOT;
- Recherche de proximité : non disponible;
- Troncature : le symbole * remplace n'importe quoi;
- Sensibilité à la casse : insensible;
- Recherche par champs : non disponible;
- Limites : non disponible;
- Mots fréquents : ils ne sont pas recherchés;
- Tri : dans 5 catégories et par pertinence;
- Avantage : un des plus larges répertoires et très récent;
- Inconvénient : parfois problèmes de recherche;

Yahoo

- Types de fichiers : html;
- Opération par défaut : AND ou OR par « intelligent default »;
- Recherche booléenne : + pour AND, – pour NOT;
- Recherche de proximité : termes entre guillemets;
- Troncature : automatique;
- Sensibilité à la casse : insensible;
- Recherche par champs : seulement par titre et URL;
- Limites : catégorie Yahoo, actualité (7 choix);

- Mots fréquents : ils ne sont pas recherchés;
- Tri : en 5 catégories;
- Avantages :
 - très populaire;
 - un des plus grands répertoires;
 - beaucoup de services d'informations et d'aides;
- Inconvénient :
 - la mise à jour n'est pas régulière;
 - pondération positive des sites commerciaux;

Les moteurs de recherche morts

- Deja.com
- Direct Hit
- Excite
- Go
- Infoseek
- NothernLight

6) Les messages d'erreur

Lorsqu'on demande une page, il peut arriver qu'un message apparaisse :

- Les messages **2XX** sont des messages de réussite;
- Les messages **3XX** sont des messages de redirection;
- Les messages **4XX** sont des erreurs de requête du navigateur;
- Les messages **5XX** sont des erreurs de réponse du serveur;

Conclusion : plus le chiffre est grand, plus on est embêté !

Les messages de réussite 2XX

- **200 OK** : tout va bien, l'information va suivre;
- **201 Created** : indication de l'URL à laquelle le document nouvellement créé devrait se trouver;
- **202 Accepted** : le serveur a accepté la requête mais ne peut pas dire si une action a été effectuée ou un fichier envoyé;
- **203 Non-Authoritative** : une partie des informations n'a pas été acheminée;
- **204 No content** : le serveur a reçu la requête mais aucune donnée n'a été renvoyée;

Les messages de redirection 3XX

- **301 Moved permanently** : le document a été déplacé définitivement vers une nouvelle URL et la redirection est automatique;
- **302 Moved temporarily** : il n'est pas certain que la nouvelle URL est définitive;
- **303 See other** : le document demandé se trouve à une autre URL et peut être récupéré;
- **304 Not Modified** : le document n'a pas été modifié depuis la dernière visite;

Les erreurs du navigateur 4XX

- **400 Bad request** : mettez vos lunettes, il y a une erreur dans l'URL;
- **401 Unauthorized** : soit vous n'avez pas accès à ce site car il est protégé par un mot de passe, soit vous avez fait une erreur dans votre mot de passe;
- **402 Payment required** : sortez votre carte de crédit, l'information est payante;
- **403 Forbidden** : c'est une zone interdite, circulez y a rien à voir;
- **404 Not Found** : la page a disparu ou elle a déménagé, recherchez la sur le site général; parfois une erreur 404 cache une erreur 400, ayez l'œil;

Les erreurs du serveur 5XX

- **500 Internal Server Error** : le serveur n'est pas en mesure de transmettre correctement l'information; réessayez plus tard;
- **501 Not Implemented** : le serveur ne sait pas répondre à la requête introduite; secouez votre informaticien réseau;
- **502 Service temporarily overloaded** : C'est l'heure de pointe, il y a un bouchon; réessayez plus tard;
- **503 Service unavailable** : le service demandé (autre que http) n'est pas retourné au serveur correctement; réessayez plus tard;

Les autres erreurs

voir <http://www.fundp.ac.be/uam/tp/erreurs.html>

7) Les aspirateurs de site

Les aspirateurs de sites sont des logiciels qui permettent de **télécharger tout un site web pour pouvoir le consulter hors connexion**. Un paramètre à préciser est la profondeur de téléchargement.

- pour PC : SmartBud (gratuit), HTTrack (gratuit), ...
(voir le site <http://www.telecharger.com> rubrique Internet – Aspirateurs de site)
- pour Mac : Internet explorer permet d'enregistrer un site jusqu'à une profondeur de 5 et d'autres logiciels existent mais très peu sont gratuits;

Les références

- ce cours est basé sur le Cours de recherche d'information sur internet de Guillemette Lauters (septembre 2002)

<http://www.fundp.ac.be/uam/tp>

- les informations sur les moteurs de recherche ont été actualisées d'après le super site suivant :

<http://www.searchengineshowdown.com>

