La Bibliothèque des Sciences et Techniques de l'ULB

L'Internet: un nouveau mode de communication et de diffusion de l'information

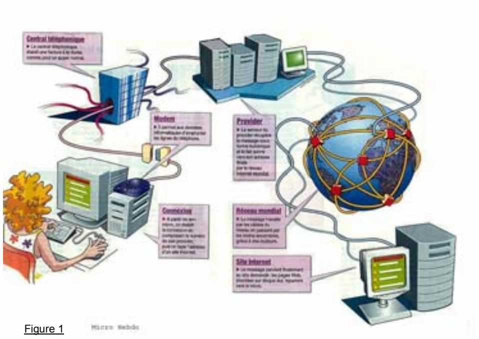


Figure 1 Micro Hebdo

Internet est un vaste réseau formé de centaines de réseaux dans le monde entier. Tous les ordinateurs connectés à ce réseau communiquent entre eux grâce à un protocole commun, le protocole TCP/IP et utilisent l'architecture client/serveur.

Les composantes d'Internet les plus utilisées:

E-Mail:

L'E-Mail ou messagerie électronique est un service qui permet d'envoyer et de recevoir des messages textuels ou des fichiers sur Internet.

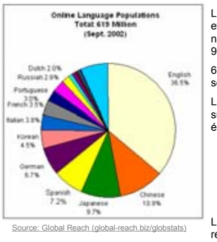
FTP:

FTP ou « File Transfert Protocol » est un protocole qui permet le chargement et le déchargement de fichiers via Internet

Wold Wide Web:

Le WWW, basé sur le principe de l'hypertexte, permet de naviguer de manière conviviale parmi les ressources du web. Le web fournit une interface unique d'accès à différents types de ressources, à des fichiers de formats divers (web, image, vidéo, etc…) ainsi qu'aux différents protocoles d'accès (FTP, Telnet, e-mail…)

Figure 2

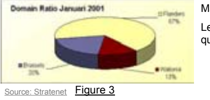


Source: Global Reach (global-reach.biz/globstats)

L'internaute moyen est un homme âgé entre 15 et 50 ans, il est majoritairement anglophone et navigue avec Internet Eplorer 5.x sous Windows 98.

67% des Américains utilisent Internet contre seulement 32% des belges.

Les activités principales sur Internet sont: surfer sur la toile: 82% et utiliser le courrier électronique: 81.5%.

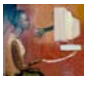


Source: Stratenet Figure 3

La Belgique compte 3,2 millions d'internautes réguliers en 2002, à savoir une augmentation de 16% par rapport à 2001 (Belgian Internet Mapping de InSites Consulting).

Le nombre de Wallons connectés reste plus bas que la moyenne nationale.

Le web: des sites de rencontre, chat, hot … mais pas uniquement



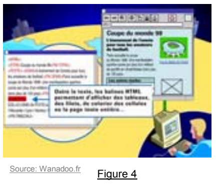
Chaque image, page, dossier sur le Web possède sa propre adresse. Cette adresse est appelée URL (Uniform Resource Locator).

Elle identifie l'endroit exact sur le réseau où se situe une ressource ainsi que le protocole nécessaire (http, ftp, telnet, etc.) pour pouvoir accéder à l'information.

L'adresse Internet se compose en général de la façon suivante:

HTML

HTML (Hyper Text Markup Langage) est un langage de description de documents qui permet de concevoir les pages web et de décrire les liens hypertextes.



Source: Wanadoo.fr Figure 4

L'HTML offre la possibilité de faire des liens à l'intérieur d'un document, vers un autre document html, vers des images, des vidéos, des fichiers audio…

Le logiciel navigateur utilise les informations contenues dans le langage HTML pour restituer à l'écran l'aspect des pages web.

L'information sur le web peut être scindée en deux catégories:

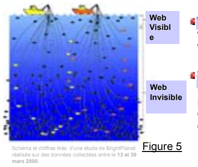


Figure 5

- Web visible ou « surface web » qui regroupe les pages web statiques
- Web invisible ou « deep web » regroupe l'ensemble des pages web générées dynamiquement à la demande de l'utilisateur.

Le web invisible représente environ 550 milliards de documents uniques soit **260 fois le web visible. L**es 60 sites les plus importants représentent à eux seuls plus de 40 fois le volume du web visible.

Ce sont les pages accessibles par mot de passe ou exigeant une autorisation préalable, celles pour lesquelles il faut remplir un formulaire pour afficher les pages, plus important encore c'est toute l'information résidant dans les bases de données traditionnelles - qu'elles soient bibliographiques ou textuelles.

Le contenu du Web Invisible émanant de professionnels est de qualité supérieure au web visible.

Le Web Invisible est la catégorie augmentant le plus rapidement sur le net.

95% du Web Invisible est accessible gratuitement sans restriction ni inscription.

Distribution de l'Information sur le web

Le Web Visible

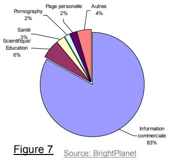


Figure 7 Source: BrightPlanet

Les pages du Web Visible sont principalement des pages commerciales.

Vu la facilité de publication sur le Web Visible et le fait qu'aucun contrôle de qualité ne s'applique à ces pages, le contenu informationnel est très pauvre.

Par exemple: pages personnelles, photos de vacances, pages publicitaires…

Le Web Invisible

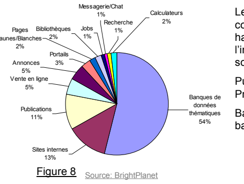


Figure 8 Source: BrightPlanet

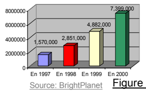
Le contenu du Web Invisible est principalement constitué de bases de données scientifiques à haute valeur ajoutée, le plus souvent l'information est validée par des comités de scientifiques.

Publications d'articles scientifiques en ligne: Pre-print Loos Alamos en Physique…

Bases de données scientifiques: PubMed, la base de données biomédicales…

Trouver ce que l'on cherche et même ce que l'on ne cherchait pas…

Evolution du nb de sites web



Source: BrightPlanet Figure 10

Le nombre de sites web étant en croissance continue, il devient de plus en plus difficile de trouver l'information pertinente sur Internet, information qui est le plus souvent diluée parmi une quantité importante d'informations inutiles et non désirées.

Les outils permettant d'effectuer des recherches sur Internet sont apparus en grand nombre pour répondre à cette demande. Parmi ces outils on peut discerner les services suivants:

- Les annuaires de recherche:
  Sites présentant des listes de sites classés de façon thématique

  Points forts:
  - simple d'utilisation
  - référencement des sites incontournables
  - classement parmi des catégories structurées

  Points faibles:
  - non exhaustif
  - mise à jour souvent lente

- Les moteurs de recherche:
  Outil basé sur l'utilisation d'un robot (programme) qui parcourt toutes les pages et les indexe dans une base de données

  Points forts:
  - relativement exhaustif
  - recherche par mot-clés
  - différentes options de recherche (images,…)

  Points faibles:
  - nombre de réponses souvent très élevé
  - beaucoup de parasites

- Les méta-moteurs:
  Outils permettant de consulter différents moteurs de recherche en une seule interrogation.

Principes de fonctionnement des moteurs de recherche

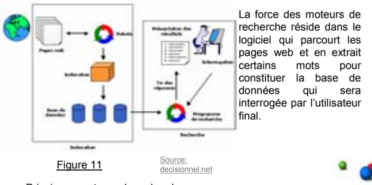


Figure 11 Source: decisionnel.net

La force des moteurs de recherche réside dans le logiciel qui parcourt les pages web et en extrait certains mots pour constituer la base de données qui sera interrogée par l'utilisateur final.

Principaux moteurs de recherche:

GOOGLE: http://www.google.com	MSN Search: http://search.msn.com	HotBot: http://www.hotbot.com
AllTheWeb.com (FAST) http://www.alltheweb.com	Lycos: http://www.lycos.com	AltaVista: http://altavista.com

Annuaires/répertoires:

Yahoo: http://www.yahoo.com	Chemdex: http://www.chemdex.org/ répertoire de sites en chimie
Infobel: http://www.infobel.com/belgium/ l'annuaire téléphonique belge	

Meta-moteur de recherche:

Profusion: http://www.profusion.com	**Dogpile**: http://www.dogpile.com	**MetaCrawler**: http://www.metacrawler.com
Fazzle (SearchOnline.info) http://www.searchonline.info/	**Excite**: http://www.excite.com	**Webcrawler**: http://www.webcrawler.com

Taille des index

GG=Google, FAST=FAST, AV=AltaVista, INK=Inktomi, NL=Northern Light

Source: searchenginewatch.com Decembre, 2001 Figure 12

Taux d'utilisation des moteurs de recherche

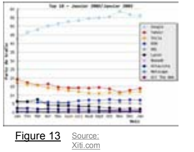


Figure 13 Source: Xiti.com

Google est actuellement le moteur de recherche le plus complet et le plus utilisé en Europe et dans le monde. 73% des pages web repértoriées sont indexées, 1.75% de la base de données concerne des formats autres que le format html (pdf, gif…), 25% des pages ne sont pas encore indexées et seulement 0.15% de la base de données est réindexée journalièrement.

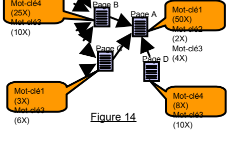


Figure 14

Google classe les sites par ordre de pertinence par rapport à une requête exprimée par l'utilisateur, ce classement est effectué grâce à un algorithme complexe appelé PageRank qui fait intervenir le nombre de citations reçues par cette page et procède à l'analyse du contenu de la page qui contient le lien pour juger de l'importance de celle-ci et les mot-clés qu'elle contient.

Poster réalisé par Dominique Lerinckx