**Les ressources Internet**

**Introduction**

En 1969, le réseau Arpanet de l'Advanced Research Project Agency (ARPA) des USA avait été lancé à cause du besoin de relier les ordinateurs entre eux par un réseau insensible aux destructions qui pourraient résulter d'une guerre nucléaire.
Tout au long des années 70, les laboratoires de recherche sont progressivement reliés au réseau Arpanet. Contrairement aux réseaux conventionnels qui ne pouvaient accommoder que des machines de même type, Arpanet présentait un énorme avantage, celui de pouvoir mettre en liaison des machines hétérogènes (Unix, IBM,...) , à condition toutefois qu'elles communiquent toutes selon le même mode. Cette normalisation impliquait une sorte de « charte », un protocole standard universel. Le protocole utilisé sur Arpanet au début des années 1970 s'appelait le NCP (Network Control Protocol). Cependant, dès 1973 on envisage l'établissement d'un nouveau protocole plus performant, capable de mieux acheminer les messages en cas de destruction partielle du réseau en cas de guerre.
Environ dix ans plus tard en 1983, le NCP est définitivement abandonné au profit du TCP/IP que nous utilisons encore aujourd'hui et qui constitue le protocole de base d'Internet. TCP est chargé de segmenter le message en paquets et de réarranger ces derniers à la réception, tandis que IP est chargé d'assurer l'acheminement des paquets d'ordinateur en ordinateur jusqu'à destination.
En 1977, Internet (INTERconnected NETworks) devient réalité car le TCP/IP est effectivement utilisé pour relier divers réseaux à Arpanet. Le nombre d'ordinateurs connectés dépasse alors la centaine et depuis, ce nombre ne cesse d'augmenter d'année en année. Internet est un réseau informatique mondial constitué d'un ensemble de réseaux d'ordinateurs qui coopèrent dans le but d'offrir une interface unique à leurs utilisateurs.
L'origine du Web remonte à mars 1989 lorsque Tim Berners-Lee, un informaticien au CERN, propose de créer sur le site Internet du [CERN](http://public.web.cern.ch/public/Welcome-fr.html) un ensemble de documents reliés les uns aux autres par des liens hypertextes, afin de faciliter la recherche d'informations pour les physiciens des particules. En 1991 (300 000 ordinateurs connectés) , alors que le nombre d'ordinateurs connectés atteint les 300 000, le plus grand site Internet au monde est celui du CERN, futur berceau du Web, un vaste ensemble mondial de documents dits hypertextes et hypermédias distribués sur Internet.
En 1992 (1 000 000 de machines connectées), le premier lien vers le site Web du CERN est créé sur le serveur de Fermilab aux États-Unis: la toile mondiale commence à se tisser. Les années 1990 ont vu le Net continuer à grossir à une vitesse exponentielle sous l'impulsion du Web. En 1993, on comptait 600 sites, en 1995 plus de 15 000, le World Wide Web devient le service le plus important sur Internet.

**C'est quoi Internet ?**

L'Internet est un réseau de réseaux reliant des ordinateurs qui partagent les protocoles TCP/IP. L'Internet est le véhicule de transport des renseignements entreposés dans les dossiers ou les fichiers sur un autre ordinateur. Il est quelquefois comparé à un système international géant de plomberie. L'Internet lui-même ne contient pas d'informations. C'est une déclaration légère erronée que de dire "un document a été trouvé sur l'Internet ". Ce serait plus correct de dire a été trouvé à travers ou en utilisant Internet. Il a, en fait, été trouvé dans un des ordinateurs connectés à Internet.

**C'est quoi le Web ?**

Le WWW incorpore tous les services Internet ci-dessus et beaucoup plus. Vous pouvez retrouver des documents, voir des images, animations et vidéos, écouter des fichiers audio, parler et écouter des voix, à condition que votre ordinateur a le matériel et logiciel pour faire ces choses.
Quand vous accédez à Internet en utilisant un navigateur web (par exemple, [Internet Explorer](http://www.microsoft.com/france/windows/products/winfamily/ie/default.mspx), [Firefox](http://www.firefox.fr/), [Mozilla](http://frenchmozilla.sourceforge.net/), Netscape, Opéra, Safari), vous visualisez des documents sur le World Wide Web. Le fondement de base selon lequel le WWW fonctionne est le langage de programmation appelé HTML. C'est HTML et les autres programmations imbriquées dans HTML qui rend l'hypertexte possible. L'hypertexte est la capacité des pages web de contenir des liens qui sont des zones dans une page ou boutons ou graphiques sur lesquels vous pouvez cliquer avec votre souris pour retrouver un autre document dans votre ordinateur. Cette fonctionnalité qui utilise des liens Hypertexte est la caractéristique unique et révolutionnaire du Web.
Comment fonctionnent les liens hypertexte ? Chaque document ou fichier ou site ou film ou fichier audio ou n'importe quoi que vous trouvez sur le Web a un URL unique (Uniform Resource Locator) qui identifie l'ordinateur dans lequel il se trouve, son emplacement dans cet ordinateur, et son nom de fichier spécifique. Chaque lien Hypertexte sur chaque page web dans le monde contient un URL. Quand on clique sur un lien sur une page Web, on envoie une requête pour retrouver le document unique sur un ordinateur dans le monde qui est identifié par cet URL uniquement. Les URLs sont comme des adresses de pages web. Un ensemble de standards internationalement acceptés (tels que TCP/IP et HTML) rendent possible ce

 phénomène du recouvrement des renseignements global qui transcende toutes frontières politiques et linguistiques.
Toutes les sources d'information convergent progressivement vers Internet. Il est ainsi possible d'obtenir des informations issues de sources formelles (catalogues de bibliothèques ou de centre de documentation spécialisé, presse en ligne, rapports, études, banques de données, brevets...) ou des sources informelles, à partir du moment où l'on réussit à se constituer son réseau personnel d'experts.
Le réseau Internet est d'abord une mémoire ou plus exactement la possibilité d'accéder à des millions de mémoires réparties sur l'ensemble du globe. L'information est récupérée sous forme électronique. Elle est donc classable, exploitable, archivable par des moyens informatiques.
Cependant Internet est un réseau à expansion rapide qui ressemble de plus en plus à un véritable labyrinthe à échelle planétaire. Un des grands problèmes actuels est de concevoir et de mettre à disposition des outils de recherche d'information qui donnent un accès rapide et pertinent à l'information.

Les ordinateurs sur Internet peuvent utiliser un ou tous les services suivants:

* Le courrier électronique («email») aussi appelé courriel permet à un individu de communiquer électroniquement avec une personne ou un groupe de personnes. C'est actuellement le service le plus utilisé dans Internet.
* Les listes d'envoi («listserv») sont des listes de distribution permettant d'acheminer un message qui sera retransmis à tous les individus abonnés à cette liste. On peut s'abonner ou se retirer de ces listes d'envoi qui sont souvent spécialisées et limitées à un nombre restreint d'individus.
* Les forums électroniques (Usenet) aussi appelés babillards électroniques favorisent les échanges entre de nombreuses personnes puisque les messages sur les sujets les plus divers sont envoyés à tous les serveurs Usenet du monde où ils pourront ensuite être consultés.
* Telnet est un protocole permettant à un ordinateur de se brancher à un autre ordinateur comme s'il était un simple terminal. Ce service tend à disparaître puisqu'on le retrouvait surtout autour des catalogues de bibliothèques qui sont maintenant de plus en plus disponibles en format Web.
* FTP («File transfer Protocol») est un protocole qui permet d'échanger entre ordinateurs des fichiers contenant des programmes, des textes, des images, etc.
* Le «chat» est un mode de communication en direct où l'on dialogue par écrit. Les sites par le biais desquels on peut faire du bavardage-clavier sont en général reliés à des thèmes ou à des intérêts spécifiques. On peut les repérer grâce à des moteurs de recherche.
* La téléphonie est un service moins connu et utilisé. Il permet de converser en direct pourvu que les utilisateurs aient à leur disposition carte de son, microphone et haut-parleurs.
* Les blogues sont des journaux personnels en ligne, tenus au moyen de logiciels simples qui permettent de saisir un texte sur un ordinateur et, une fois connecté, de l'envoyer instantanément pour qu'il s'affiche sur une page Web entretenue à cet effet. Souvent considérés comme une version améliorée des anciennes «pages perso» sur le net, les blogues ne se bornent pas à cette fonction. Ils permettent davantage : graphisme plus intéressant, archivage automatique des données, possibilité d'auteurs multiples. Les blogues mélangent volontiers informations et opinions, et sont souvent accompagnés de liens à des sources originales, à d'autres blogues ou à des articles que le blogeur commente ou signale à son public. Le premier blogue reconnu comme tel remonte à 1994.
* Les lecteurs RSS, interprétant XML de manière analogue avec laquelle les navigateurs interprètent le HTML.

**Typologie des informations sur Internet**

Internet représente une source d'information très importante. Il est important de connaître les différents types d'information disponibles afin d'être plus efficace dans ses recherches.
Les millions de pages disponibles sur le réseau Internet sont de nature et de forme trop différentes pour constituer des "documents" au sens traditionnel du terme.
On peut trouver toutes sortes de textes (livres, articles, dictionnaires...), de nombreux instruments de référence (catalogues de bibliothèques, banques de données, répertoires..), des informations d'actualités politiques, sociales, économiques, culturelles (dossiers, journaux, dépêches..), des données chiffrées et statistiques, des renseignements sur des services administratifs, des images (photos, tableaux, animations, ...) et du son (émissions de radio, musique..), des programmes,...
Chaque type de document est numérisé sous un format spécifique qui nécessite, pour le visualiser ou le retraiter, des logiciels particuliers. Les navigateurs récents intègrent désormais de nombreux "plug-ins" permettant de visualiser certains documents.

Parmi ces informations nous retrouvons :

* Les informations économiques : De nombreuses informations économiques et financières sont disponibles sur internet. Les sites d'entreprises sont un bon point de départ. De plus en plus d'entreprises possèdent un site Internet, véritable vitrine, permettant à ces dernières de présenter leurs produits, et de fournir des informations (sites d'entreprise). Des sites plus généralistes comme ceux de la presse économique sont des passages obligés pour les gens à la recherche d'informations économiques.Exemples : [Maroc Telecom](http://www.iam.net.ma/), [L'économiste](http://www.leconomiste.com/)
* Les organisations : Les organisations mondiales (le domaine est généralement en .org) représentent également une source d'information importante. Les informations dispensées sont fiables.Exemples : [OCP](http://www.ocpgroup.ma/), [FAO](http://www.fao.org/)
* Les sites gouvernementaux : Ils offrent des informations pratiques, fiables. Vous pouvez consulter des informations sur un domaine précis ou bien télécharger des rapports ou des documents officiels. Exemples :[Ministère de l'agriculture](http://www.agriculture.ovh.org/madrp.htm), [Ministère de l'aménagement du territoire](http://www.matee.gov.ma/), [Administration des douanes](http://www.douane.gov.ma/)
* La presse, les médias : Le nombre de quotidiens, d'hebdomadaires, de radio, de télévisions qui possèdent un site web va grandissant. Ces sites offrent la possibilité de consulter des informations du type actualités en ligne. la mise à jour est fréquente et l'information est fiable. Exemples : [TELQUEL](http://www.telquel-online.com/), [Le Monde](http://www.lemonde.fr/), [TF1](http://www.tf1.fr/), [2 M](http://www.2m.tv/)
* Les pages personnelles : Internet offre la possibilité à tout le monde de pouvoir s'exprimer sur n'importe quel sujet. Les pages personnelles souvent hébergées gratuitement ([chez.com](http://www.chez.com/), [geocities.com](http://www.geocities.com/)...) traitent de tout... et parfois n'importe quoi. Elles peuvent être intéressantes pour trouver des avis sur un sujet donné. Cependant la fiabilité de ces pages n'est pas toujours à toute épreuve. Exemples : [CineCritic](http://www.multimania.com/critic/), [Surf](http://www.chez.com/poomcat), [Plongée](http://www.chez.com/limouz)
* Les Universités. : Au delà de présentations générales, vous pouvez maintenant accéder à des informations plus nombreuses (publications des professeurs, compte-rendu de recherche...). Exemples : [IAV Hassan II](http://www.iav.ac.ma/), [Dauphine](http://www.dauphine.fr/), [Al Akhawayn University in Ifrane](http://www.aui.ma/)
* Informations pratiques : Que vous souhaitiez préparer un déplacement, un voyage, connaitre la météo,... Internet peut vous aider à tous moments. Exemples : [tanmia](http://www.tanmia.ma/), [Royal Air Maroc](http://www.royalairmaroc.com/),[Itinéraire](http://www.iti.fr/), [Météo](http://www.meteo.fr/)
* Les outils : Utiliser un dictionnaire, traduire un texte, convertir des devises, Internet est là pour vous aider. Exemples : [Dictionnaire](http://eurodic.ip.lu/cgi-bin/edicbin/EuroDicWWW.pl), [Devises](http://www.xe.net/ucc/), [Traducteur](http://trans.voila.fr/)
* Les bibliothèques : De nombreuses bibliothèques se sont ouvertes aux nouvelles technologies. Elles offrent la consultation de leur catalogue mais aussi des documents en ligne. Les librairies en ligne donnent au livre une seconde jeunesse. Ces outils sont indispensables pour les recherches. Exemples : [Bibliothèque Nationale](http://www.bnf.fr/), [Bibliothèque Nationale du Royaume du Maroc](http://www.bnrm.ma/fr/index.htm) ,[Mohammed VI Al Akhawayn University Library](http://www.aui.ma/library/)

**Caractéristiques de l'information sur Internet**

* Abondante : Pour une requête qu'on formule, il est souvent qu'on obtienne une quantité trop importante d'information. Bien qu'abondante, souvent cette information est souvent trop abondante et peu pertinente. Aucun des moteurs de recherche existants n'est suffisamment ne fournit des résultats réellement ciblés et en adéquation avec nos besoins.
* Accès difficile: La recherche d'informations n'est pas toujours très facile sur Internet car aucun standard ni aucune réglementation ne sont imposés aux fournisseurs d'information. Les différents annuaires et moteurs de recherche ont leurs propres règles. Il est On peut ne trouver le réponse à notre requête qu'après avoir passé en revue plusieurs moteurs de recherche. Parfois il suffit de formuler à nouveau la requête pour obtenir de meilleurs résultats.
* L'information se renouvelle sans cesse : Internet est le média où l'écrit se renouvelle le plus. L'extrême facilité de publication fait qu'une information peut apparaître à tout moment dans n'importe quelle sphère d'Internet, puis disparaître. La première difficulté consiste donc à pouvoir surveiller au plus près les évolutions dans un domaine donné.
* L'information n'est pas toujours fiable : Aucune fonction de validation de l'information par un comité éditorial n'est assurée. Le bon peut côtoyer le pire et l'information pertinente être masquée par la publicité ou la désinformation, volontaire ou non.

**Le Web invisible**

Les meilleurs moteurs de recherche indexent moins de 20% du Web. En effet, les outils de recherche (en particulier les moteurs) ne référencient pas les pages suivantes :

* Les fichiers qui ne sont pas au format HTML (certains moteurs comme Google indexent cependant d'autres formats comme le pdf, les documents MSOffice,...)
* Les pages Web demandant un login ou restreintes à un domaine particulier
* Les pages dont le contenu est dynamique (créées en réponse à la demande d'un utilisateur par exemple)
* Les sites intranet
* Les pages qui ne sont pas référenciées (aucun lien ne pointe vers cette page)
* Les sites qui utilisent un fichier robot.txt interdisant au moteur d'indexer les pages
* Les ressources ne provenant pas du Web (sites FTP, Gopher,...)

Les documents constituant le Web invisible peuvent être dynamiques (non localisables), non référencés (volontairement ou non), ou de nature non indexable (ex. les animations Flash …). En voici les principales raisons :

* Sites protégés par mots de passe et Intranet.
* Problème des cadres : une structure de cadres (frames) se compose d'une page mère et de pages filles. Le contenu informatif se trouve toujours dans les pages filles. La difficulté est que lorsqu'une page mère est soumise aux moteurs, ceux-ci ne sont souvent pas capables de retrouver les pages filles et de les indexer. Seule la page mère se retrouve alors indexée. Mais comme il s'agit d'une page sans contenu, le référencement est inefficace.
* Formats de fichiers : HTML est le format du web reconnu par tous les robots mais beaucoup se sont ajoutés. Seul Google indexe d'autres formats pdf, Microsoft Excel (xls), PowerPoint (ppt), Word (doc), et Rich Text Format (rtf), ...
* Pages interdites de référencement en utilisant un fichier Robots.txt : Cette interdiction est posée par l'auteur des pages ou le gestionnaire du site.
* Pages HTML statiques/pages HTML dynamiques :
* Page HTML statique : Fichier ASCII qui a été structuré à l'aide de balises HTML, dans le but de le diffuser dans le Web. Le fichier reste tel quel tant qu'il n'est pas modifié ou détruit par son auteur ;
* Page HTML dynamique : Page HTML créée en réponse à la demande d'un utilisateur, dont la forme est fixe et le contenu variable est issu d'une base de données, ce qui permet ainsi de l'adapter aux critères de recherche de celui-ci.

La difficulté est que les robots ont besoin de suivre des liens à l'intérieur du site afin de pouvoir indexer les pages correspondantes. Si les pages dynamiques ne peuvent être accédées que par le biais de l'interrogation d'un formulaire, celles-ci ne pourront pas être indexées par les moteurs. Il faut donc intégrer, dès les premières pages, des liens vers les principales Url dynamiques que l'on souhaite voir indexées. Ceci peut se faire de façon invisible, de manière à ne pas remettre en cause le design et la structure du site.

* Pages tronquées : Certains robots n'indexent pas entièrement les pages. Cette information est rarement publique.
* Les robots stoppent quand ils rencontrent une URL contenant un point d'interrogation.
* L'absence de liens hypertextes a été mise en évidence par l'étude d'IBM connue sous le nom de la théorie du n'oeud papillon d'IBM.

Des chercheurs d'IBM, Compacq et Altavista ont dressé une cartographie de l'Internet à partir de l'analyse des liens établis entre pages web. La structure du réseau que l'on croyait comparable à celle d'une toile d'araignée, présente en réalité 4 ensembles de pages organisés en noeud papillon. On distingue le coeur du réseau très interconnecté, puis un deuxième ensemble de pages ayant établi des liens vers ces dernières, enfin un troisième groupe facilement repérable depuis le noyau. Les autres pages, quant à elles, restent déconnectées et isolées du réseau.

* L'information de presse : de plus en plus de sites de presse (quotidiens, radio, télévision…) diffusent en temps réel de l'information. Les robots classiques sont incapables de suivre.