

WORLD WIDE WEB : LE TELEPHONE RUSSE

L'Internet s'étend tous les jours aussi bien du point de vue du nombre de serveurs connectés que du point de vue de la quantité d'information disponible sur ces serveurs et du nombre de clients qui se connectent. Les principaux services responsables de cet engouement sont le World Wide Web et l'e-mail. Le 3W, le WWW, le World Wide Web ou en un mot le Web qui désigne en anglais la toile d'araignée représente donc la toile d'araignée couvrant le monde entier. Comme on le sait, ses concepteurs ont eu l'intelligence d'englober les principaux outils déjà existant à l'époque où il fut inventé (ex: Gopher, les News, ftp, telnet et d'autres) ce qui a enlevé tous les freins à son acceptation et a largement contribué à amorcer l'explosion de son utilisation. Surfer sur le Web c'est visiter d'autres pays à l'autre bout du monde, ou une galerie de peinture dans votre ville, avoir les dernières nouvelles ou accéder à des traités sur la préhistoire, réviser votre Anglais ou apprendre le Breton, réserver une place de ciné ou gérer votre compte en banque... et la liste est longue. Mais au delà de l'activité de consultation la force du Web et le secret de son expansion sont le fait que tout le monde peut y participer. Tout le monde peut créer sa ou ses pages Web et mettre sa petite touche dans cette toile sans maître qu'est le Web. C'est pour cela que ce chapitre s'intéresse à la technologie sous-jacente au Web.

I. LES BASES DU WEB

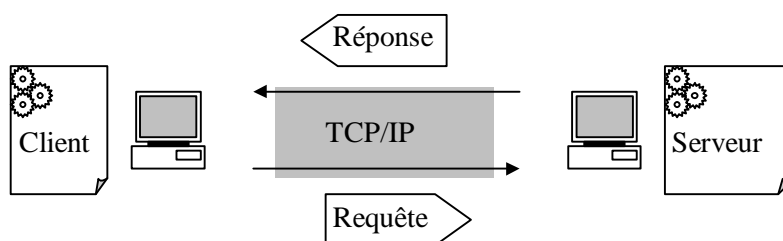
Le Web fait appel à un certain nombre de concepts de base. Nous verrons les trois plus importants dans cette première partie.

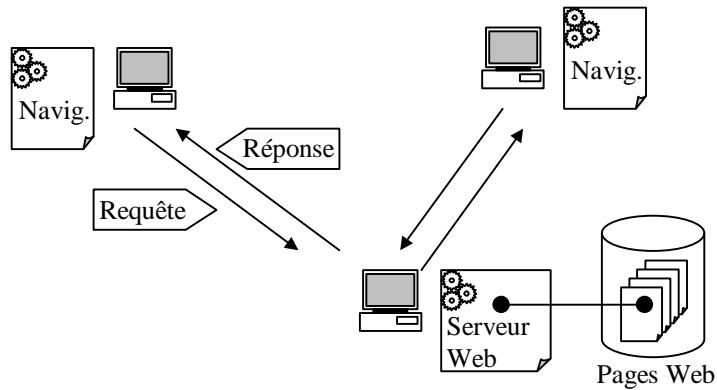
I.1 ARCHITECTURE CLIENT-SERVEUR

Le World Wide Web s'appuie sur la notion d'architecture client serveur. Un serveur est une machine en général assez puissante qui fournit un ou plusieurs services (accès à des sources de données, applications...) . Pour fournir ces services elle fait tourner en permanence des programmes que l'on appelle aussi des serveurs en l'occurrence ce sont des serveurs Web ou serveurs HTTP. De l'autre coté les utilisateurs font tourner sur leur machine (machine cliente) un programme client qui, comme son nom l'indique va être demandeur de services, en l'occurrence ce client est un navigateur Web qui va demander des pages Web à un serveur Web. Le dialogue entre le client et le serveur se compose donc de requêtes émises par le client et de réponses données par le serveur.

Le client est couramment appelé un **navigateur** (ou encore browser, fureteur ou butineur). Les navigateurs les plus connus étant Netscape, Internet Explorer, Lynx, Mosaic, Opera, Kfm. Les plus courants acceptent des extensions (Plug-In) permettant d'étendre leurs capacités (lire des vidéos, recevoir du son ou des films en flot continu, ...). Ils connaissent aussi fréquemment des langages évolués (JAVA, Javascript, VRML...) permettent d'élargir le champ des possibilités de l'utilisation des pages Web. Les navigateurs sont des logiciels soigneusement étudiés pour faciliter et assister la navigation sur le Web. Ils proposent en standard des fonctionnalités d'historique pour revenir sur ses pas, une gestion des signets pour pouvoir garder ses pages préférées bien organisées et facilement accessibles, et plusieurs autres fonctions utilitaires et outils d'assistance.

Un **serveur** est une machine qui est capable de 'servir' d'autres machines en fonction de leur requête, ces dernières sont appelées 'clients'. Pour cela elle doit toujours être connectée au réseau et exécuter le démon (daemon) correspondant au service rendu. On appelle démon un programme qui tourne en tâche de fond sur une machine et le cas échéant répond à des requêtes qui lui sont adressées ou déclenche des actions en réponse à des événements ou un planning. Sur le Web les documents s'échangent selon le protocole HTTP (HyperText Transfer Protocol) et le démon qui se charge de répondre aux requêtes des autres machines se nomme HTTPD (HyperText Transfer Protocol Daemon). De la même façon un serveur offrant des fichiers via FTP est une machine sur laquelle tourne un serveur FTP encore appelé démon FTPD (File Transfer Protocol Daemon), de même pour l'e-mail, etc...





Il y a deux cas : soit l'utilisateur cherche à visualiser une page disponible sur sa machine auquel cas le navigateur obtient le fichier par simple lecture directe sur un disque de la machine sur laquelle il s'exécute, soit l'utilisateur souhaite accéder à une page disponible sur une machine distante auquel cas le navigateur doit se connecter au serveur publiant cette page à travers le réseau. On se rappelle qu'Internet est l'infrastructure internet (interconnected networks = réseaux interconnectés) d'un réseau informatique mondial. Ce réseau mondial se compose de réseaux d'ordinateurs locaux interconnectés et dont les échanges suivent les protocoles TCP, UDP et IP (Transmission Control Protocol and Internet Protocol) chaque ordinateur connecté étant adressé par un numéro IP ou un nom symbolique / nom de domaine. L'architecture client-serveur du Web repose sur ces bases en ce sens que le programme client (navigateur) se connecte au programme serveur (serveur Web) grâce aux protocoles TCP/IP et ainsi met en place une connexion bidirectionnelle fiable qu'il va utiliser pour obtenir les informations souhaitées (document, image et autres fichiers).

I.2 URL

Pour accéder à une page web il faut d'abord pouvoir décrire où elle se trouve. Pour repérer un document, un fichier, une source de données ... on a développé la notation URL (Universal/Uniform Resource Locator). Un URL peut désigner un serveur ftp, un fichier sur votre disque, un serveur gopher, une image, une adresse courrier, un serveur de News, un serveur telnet et bien sûr une page Web publiée par un serveur http, c'est-à-dire un serveur de Web. En particulier, dans ce dernier cas l'URL contient le nom du protocole d'accès au fichier (HTTP, SHTTP), le nom du serveur (adresse IP ou nom symbolique), le chemin d'accès au fichier et bien sûr le nom du fichier :

<Protocole> ://<nom serveur>/<chemin>

Exemple:

http://www-sop.inria.fr/acacia/personnel/Fabien.Gandon/index.html

Un des problèmes posés par ce système est que si un URL vient à changer il faut remettre à jour tous les liens qui l'utilisent. Il faut donc maintenir ses pages et régulièrement vérifier que les liens sont toujours corrects ou sinon les surfeurs risquent de se retrouver dans une impasse avec l'écran un message d'erreur du style "Ce document n'existe pas (erreur numéro 404)"

Les noms d'URL utilisent les lettres de l'alphabet en général en minuscule, les chiffres sont autorisés, certains caractères / . : # ont une signification particulière et sont donc réservés, enfin certains caractères sont dit non sûrs dans la mesure où ils sont interprétés ou interprétables différemment : les blancs, les étoiles, etc.

Les trois caractères / . : sont des **séparateurs** simples, le ? est un séparateur introduisant une requête qui en général demande au serveur d'exécuter un programme (CGI,...) pour générer la réponse. Il exemple typique est celui des moteurs de recherche où ce type d'URL est utilisé pour envoyer vos mots clefs à un programme qui génère la page des réponses. Exemple: si vous lancez une requête sur le mot clef 'vin' sous Yahoo! l'URL est :

http://fr.search.yahoo.com/search/fr?p=vin

Cela signifie que votre requête appelle un programme de recherche dans l'annuaire de Yahoo! avec le paramètre 'vin'.

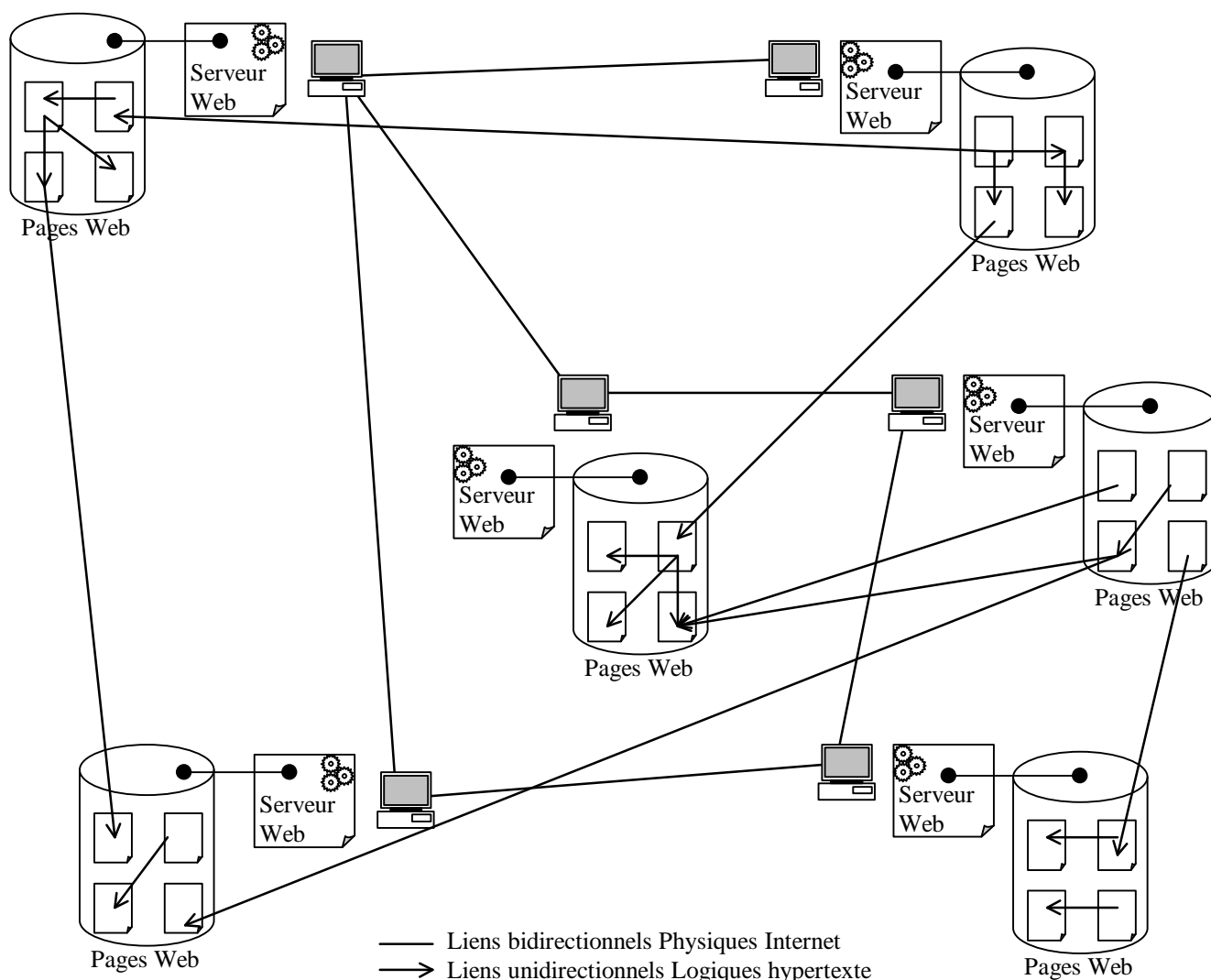
On trouve d'autres types d'URL, chacun représentant un service donné, un certain nombre d'exemples vous sont donnés dans le tableau ci-dessous la forme la plus complexe d'un URL étant :

<service>:[//][nom utilisateur]:[mot de passe][@]<serveur>:<port>/<chemin>

| Service | Masque | Exemple |
|----------------------|---------------------------------------------------|----------------------------------------|
| Web | http://serveur:port/repertoire/fichier.html | http://www.chez.com/toto/fichier.html |
| FTP | ftp://serveur/repertoire/fichier | ftp://inria.ftp.fr |
| Fichier local sur PC | file:///disque /repertoire/fichier | file:///c /tmp/fichier.txt |
| Mail | mailto:nom@organisation.domaine | mailto:Fabien.Gandon@sophia.inria.fr |
| Telnet | telnet://Nom:Password@serveur:port | telnet://gandonf:abcde@gopa.insa.fr:23 |
| Gopher | gopher://serveur:port/repertoire/fichier#marqueur | |
| Serveur de News | newsr://serveur:port/repertoire/nom.de.la.news | |
| WAIS | wais://<host>:<port>/<database> | |

I.3 HYPERTEXTE ET TOPOLOGIE DU WEB

Le World Wide Web est un vaste ensemble de sources d'informations accessibles à travers le réseau Internet. Nous l'avons vu dans l'historique, il fut initialement construit par le CERN pour la documentation des projets de recherches. Il est maintenant utilisé par tout le monde pour mettre en ligne (i.e. rendre accessible sur le Web via Internet) des documents et des services de tous horizons. L'information est présentée essentiellement sous forme de texte et d'images, mais le son, la vidéo... bref le multimédia étant en pleine explosion on commence à employer de plus en plus souvent le terme hypermédia au lieu d'hypertexte. On qualifie d'**Hypertexte** (terme et notion inventés par Ted Nelson en 1960) un document essentiellement textuel, dynamique, capable de changer et de réagir en fonction de certains événements comme par exemple un clic à la souris. Un tel document offre une très grande convivialité et la tendance étant à élargir cette technique à d'autres médias (image, video, animations...) on parle maintenant d'hypermédia. Le langage permettant de décrire les pages Web est le HTML (Hyper Text Markup Language). Ce langage à balise permet de doter certains mots, ou images d'une propriété d'hyperlien ou plus simplement de **lien** qui est constitué d'une adresse URL que vous atteindrez en cliquant dessus.



L'information disponible sur le Web a cette caractéristique qu'elle est distribuée sur une zone géographique très grande et au sein d'une même page web peuvent être conjuguées des ressources placées aux quatre coins de la planète. Elle s'organise en pages mises à dispositions sur les serveurs. Une Page Web contient donc du texte, des images... et des liens vers d'autres pages Web ou d'autres fichiers. Les liens permettent de naviguer de pages en pages d'un simple clic. L'utilisateur peut passer en un clic d'une page placée sur un serveur à San Francisco à une autre sur un serveur à Tokyo.

Le Web tire son nom du fait que, de par les liens que l'on tisse entre les pages, on construit une **toile d'araignée** gigantesque (mais pas symétrique à la différence de son homologue naturelle) qui croit et évolue de façon complètement arbitraire. Ce fouillis d'interconnexion permet entre autres choses au Web d'assurer plusieurs chemins vers la même information.

En résumé, le World Wide Web est une toile d'araignée de serveurs d'informations reliés les uns aux autres par des liens physiques (le réseau matériel) et des liens logiques (les liens hypertextes) entre les pages qu'ils publient. Attention **un lien hypertexte est unidirectionnel**: Si A pointe sur B, B ne pointe pas forcément sur A !!!

II. LA TECHNOLOGIE DU WEB

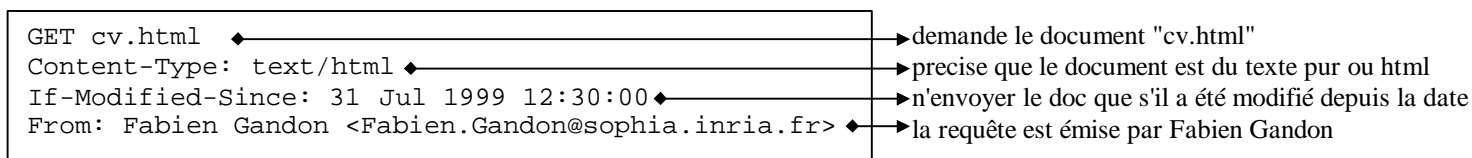
Internet en lui même n'assure pas la compatibilité et l'accessibilité de documents, ce n'est que la partie matérielle de la connexion. En revanche, le World Wide Web (WWW ou W3 ou simplement le Web) est un ensemble de protocoles (ex : HTTP) d'outils (ex : HTTPD, navigateurs...) et de normes (URL) permettant de créer, formater, rechercher échanger... bref partager de manière interactive des informations hétérogènes à travers Internet sur le principe du Client/Serveur. Le formatage des informations est principalement basé sur la technique des documents hypertextes balisés grâce au langage HTML et diffusés grâce au protocole HTTP. Mais le Web est aussi capable d'utiliser d'autres protocoles tels que : FTP (File Transfer protocol), Telnet, NNTP (Network News Transfer Protocol), WAIS (Wide Area Information System/Server), gopher (de 'go fer'), ...

II.1 HTTP

Le **protocole** de base du World Wide Web est HTTP (HyperText Transfer Protocol) qui peut être utilisé pour n'importe quelle application client-serveur impliquant de l'hypertexte. Ce protocole est capable d'assurer le transfert de texte, hypertexte, fichiers audio, images ou tout autre type d'information pouvant se mettre sous la forme d'un fichier.

Le scénario de dialogue classique entre un navigateur et un serveur Web est le suivant. Le navigateur Web client établit une connexion TCP avec le serveur Web qui contient la page qui l'intéresse. Une fois la connexion établie, le client émet une **requête** HTTP contenant une commande, une URL, et parfois d'autres informations. Lorsque le serveur Web reçoit la requête il essaie d'exécuter la commande qu'elle contient. Il retourne ensuite comme **réponse** le résultat obtenu qui peut être des données, un message d'erreur, et d'autres informations. Une fois que le client a reçu sa réponse la connexion est fermée et détruite.

Voici par exemple une requête émise par un navigateur dont l'utilisateur est "Fabien Gandon" et demandant à ce que le document html "cv.html" lui soit envoyé s'il a été modifié depuis le 31 juillet 1999 à 12:30:00.



II.2 HTML

Le langage HTML (**HyperText Markup Language**) est utilisé sur le système de partage de l'information mondial WWW (World Wide Web) depuis 1990. Ce langage se compose d'un ensemble d'annotations, appelées étiquettes ou balises, qui permettent de créer et formater un document hypertexte. Un fichier HTML est un fichier texte ce qui a l'avantage de le rendre facilement lisible sur n'importe quelle plate-forme/ordinateur. Les balises du HTML sont insérées dans le texte du document et guident son affichage. Le navigateur interprète les commandes HTML contenues dans le document et en déduit le format d'affichage du document.

HTML est le langage standard d'édition de pages hypertexte pour le Web. Il existe plusieurs versions les plus communes et les plus supportées sont les versions 2.0 et 3.2. La toute dernière étant la version 4. Une page Web peut être créée directement avec un simple éditeur de texte en tapant des commandes HTML ou en utilisant un éditeur de page Web qui très souvent vous

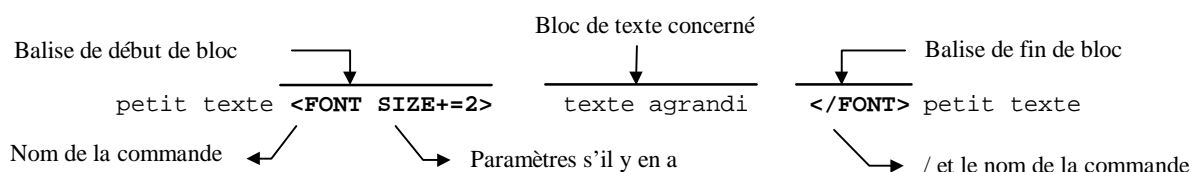
permettra de créer votre document de façon très conviviale et générera pour vous le code HTML correspondant sans que vous ayez à connaître ce langage.

Le HTML n'est pas un langage de programmation, c'est un langage d'édition de documents. Une **balise** est un mot clé, une commande du langage insérée dans le corps du document pour introduire un effet particulier (début de mise en gras, fin de mise en gras, début de tableau...). Une balise commence toujours par un signe "<" et se finit toujours par un signe ">". La plupart des balises doivent être ouvertes et fermées pour délimiter leur zone d'influence. La balise fermante contient la même commande que la balise ouvrante, mais précédée d'un caractère /

Par exemple pour mettre un texte en gras la commande est **B** (comme "Bold" en anglais qui veut dire "en gras") la balise ouvrante est **** et la balise fermante est ****. Donc si dans la phrase "Je suis étudiant à l'Université." je veux mettre le mot 'étudiant' en gras comme ceci "Je suis **étudiant** à l'Université.", le code HTML correspondant sera :

Je suis étudiant à l'Université.


Les balises peuvent utiliser des **attributs**, pour paramétrer leur effet. Prenons pour deuxième exemple de changer la taille du texte, on pourra alors taper le code suivant :



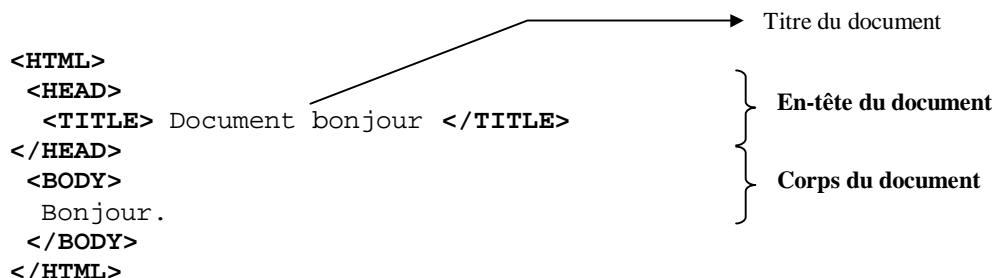
Le résultat ressemblera à : petit texte texte agrandi petit texte

Les balises disponibles en HTML vont vous permettre de formater votre document, insérer des objets (images,...) et surtout le concept central de l'hypertexte : **éditer des liens**. Un lien permet de créer une référence vers une autre page en utilisant l'URL de celle-ci. Si je reprends la phrase "Je suis étudiant à l'Université." et que je ne veux plus du mot en gras mais je veux créer un lien partant du mot 'Université' et pointant sur la page [http:// www.unice.fr/](http://www.unice.fr/), de façon à ce qu'une personne cliquant sur le mot 'Université' se voit transférée sur la page de l'Université de Nice, le code HTML correspondant sera alors :

Je suis étudiant à l'Université.

Sa réalisation à l'écran sera probablement "Je suis étudiant à l'Université." le soulignement étant par défaut la façon la plus répandue de signaler un lien. Si quelqu'un promène alors sa souris au-dessus de "Université" il la verra sûrement changer de forme par exemple une main  pour lui indiquer qu'il peut cliquer sur ce mot car c'est un lien.

Comme dans tous les langages il y a un minimum d'informations à donner dans le fichier HTML. Le fichier HTML minimum ressemble à ceci :



Les balises <HTML> et </HTML> stipulent que ce fichier texte est formaté selon le langage HTML et délimitent le contenu à interpréter.

Les balises <HEAD> et </HEAD> viennent du mot HEADER (Entête) et délimitent l'en-tête du document contenant son titre et des informations sur son contenu.

Les balises <BODY> et </BODY> délimitent le corps du document contenant le texte, son formatage les objets et les liens qu'il inclut.

Vous pouvez taper ce petit exemple dans un éditeur de texte, le sauver (par habitude, les fichiers HTML ont pour extension .html ou .htm exemple : bonjour.html ou bonjour.htm) et l'ouvrir grâce à l'option 'ouvrir un fichier' du menu fichier de votre navigateur.

Comme nous le disions au début, pour que l'édition de grands documents ne soit pas fastidieuse, on peut utiliser un éditeur HTML cependant il faut savoir que toutes les subtilités du langage ne sont pas forcément disponibles au travers d'un tel logiciel et que le code généré n'est pas toujours de bonne qualité. Enfin insistons sur le fait que le HTML n'est pas un langage de programmation en lui même et que de plus il n'y a pas de compilation car tout se fait sur le principe de l'interprétation du document tel qu'il est décrit dans le fichier texte. L'aspect programmation n'apparaît qu'au travers de l'utilisation du Java, du JavaScript, des CGI,... et ces domaines sont à eux seuls très complexes et ne seront pas abordés ici.