

# Le World Wide Web... ou encore W3

La *Toile d'araignée mondiale* des serveurs d'information, que nous appellerons plus familièrement aussi le Web, est fondée sur un ensemble de protocoles, de structures de fichiers et de syntaxes de description de contenus hypermédias.

Le WWW ou W3, a été créé en 1991 par le laboratoire des hautes énergies, au CERN à Genève. Il était initialement destiné à fournir un mode convivial pour échanger l'information dans cette communauté de chercheurs. Sa définition officielle était : *wide-area hypermedia information retrieval initiative aiming to give universal access to a large universe of documents*. Très rapidement, la technologie Web s'est répandue, d'abord dans le monde de la physique nucléaire, puis dans l'ensemble des domaines de la recherche scientifique, et enfin, avec l'ouverture d'Internet aux usages non académiques, à l'ensemble de la communauté du réseau.

Le projet W3 a donné aux utilisateurs de l'Internet un outil efficace pour accéder à une grande variété de documents de façon très simple. Grâce à des interfaces clientes conviviales, le projet W3 a changé la façon de voir et de créer l'information des utilisateurs ; il a créé le premier réseau hypermédia réparti.

*Créée pour les besoins d'un domaine scientifique, la technologie Web s'est rapidement répandue. Elle est actuellement à l'origine du développement rapide de l'Internet.*

## Hypertexte et Hypermédia

Un document hypertexte est un simple fichier texte comportant cependant des “ liens ” soit vers d'autres parties du document lui-même, soit vers d'autres documents. Dans le système W3, ces documents peuvent être localisés sur la même plate-forme (le même ordinateur) mais aussi sur d'autres plates-formes connectées au réseau Internet (celles-ci peuvent se situer n'importe où dans le monde).

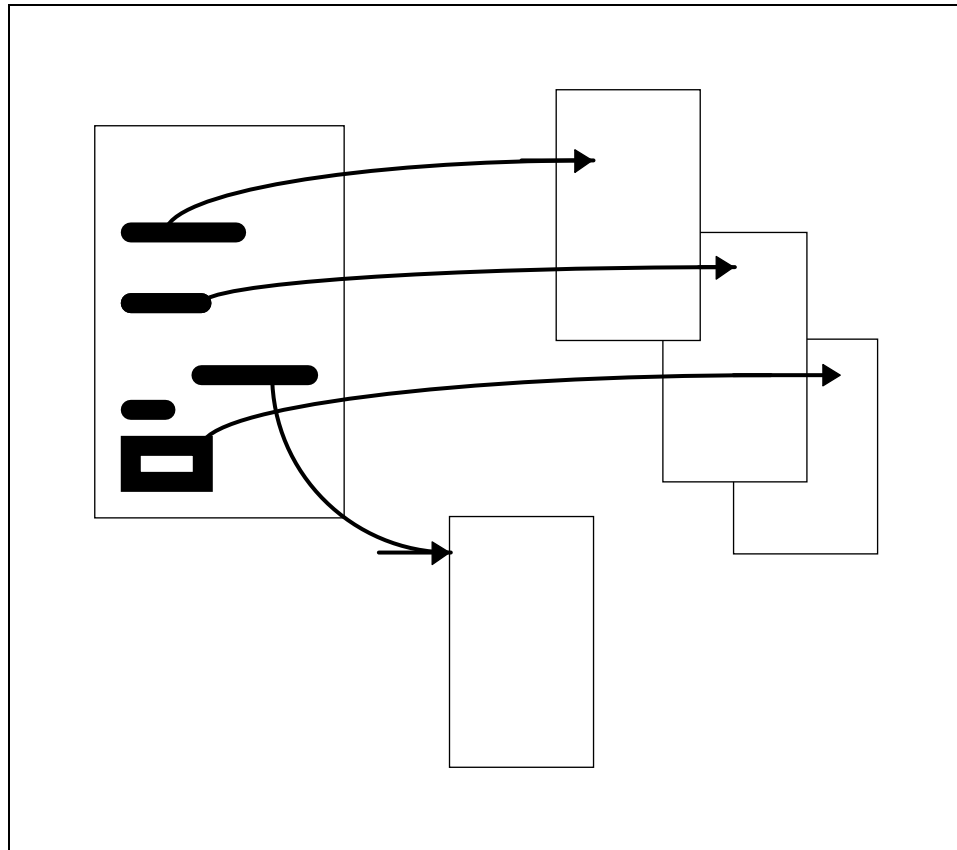
Un lien hypertexte, ou encore hyper-lien, est formé par une *ancree* et par l'*adresse* du document cible. Une ancre peut-être un mot (ou un groupe de mots) ou une image. L'ancre sera généralement mise en évidence, dans l'environnement Web, par une couleur bleue et un soulignement.

Au sein d'un document hypertexte, on peut également créer des ancres (signets) afin de pouvoir atteindre directement un point précis de ce document.

Une ancre peut donc servir soit comme origine d'un lien soit comme destination dans un document.

Un document hypermédia est un document hypertexte pouvant contenir en plus des liens vers des fichiers, images ou vidéo. Par exemple, dans un article, un clic sur le nom de l'auteur permet de chaîner sur son adresse, sa photographie, un message sonore, etc.

Le langage standard utilisé par W3 pour créer et reconnaître de tels documents hypertextes ou hypermédias s'appelle le HTML (HyperText Markup Language).



*Document hypermédia : les ancres ou boutons du texte peuvent chaîner aussi bien sur une autre « page » de texte que sur une séquence sonore, une image, une animation. De plus, le lien peut pointer soit sur un « signet » du même document, soit sur un autre document qui peut se trouver soit sur la même machine, soit sur n'importe quelle machine connectée à l'Internet dans le monde.*

## HTML (HyperText Markup Language)

HTML est le nom d'un langage informatique utilisé pour la description de documents sur les serveurs W3. Il s'agit d'un ensemble simple de commandes de formatage de documents.

La seule difficulté réside dans l'agencement de ces commandes, leur apprentissage lui-même restant très abordable.

### HTML et SGML

Le langage HTML est en fait une application de la norme SGML. SGML (*Standard General Markup Language*) est un langage de description de documents normalisé (ISO 8879, 1986), capable de décrire des documents en toutes langues, et comportant tous types d'informations.

SGML est né du besoin de disposer d'un format « pivot », capable de servir de point commun pour les échanges, mais aussi de standard intermédiaire pour les conversions de formats. En outre, le besoin de normalisation est né du développement des échanges, et en particulier de l'EDI (*échanges de documents informatisés*), qui concernent au premier chef des documents structurés.

SGML n'est pas destiné à réaliser des descriptions directes de documents, mais plutôt à décrire des structures logiques, véritables applications qui permettront, elles, la description des documents correspondants.

Parmi les applications les plus connues, on peut citer le format de description CALS (*Continuous Acquisition Lifecycle Support*) défini par le département de la Défense aux États-Unis (DoD), et de plus en plus appliqué par l'industrie, ou encore le projet DPSD de développement de normes de publication électronique, notamment pour les documentations techniques et les manuels de maintenance aéronautique.

Chacune des implémentations de SGML est définie par une structure logique (la DTD, *définition de type de documents*) capable de répondre aux besoins de codage des documents d'un groupe d'utilisateurs défini.

SGML est donc une méthode normalisée pour représenter l'information contenue dans un document, garantissant l'indépendance des systèmes de saisie et de traitement, et de la forme physique finale. Le standard définit les

principes du balisage logique généralisé, mais pas le langage de balisage, qui dépend de la classe de documents, et donc est spécifié dans la DTD.

HTML est une application de SGML, c'est-à-dire une classe de documents, définie par une DTD<sup>1</sup>. Les spécifications du langage HTML précisent en particulier qu'en cas de conflit entre HTML et SGML, c'est toujours la conformité SGML qui prévaudra.

Tous les documents HTML sont donc définis par une déclaration SGML contenant notamment la définition des caractères utilisés (jeu de caractères), ainsi que les différents types de mise en forme de paragraphes et de caractères, et les différents objets associés, tels que images, formulaires, etc. Cette déclaration, commune à tous les documents HTML, n'a pas à être transmise ; elle est simplement référencée dans l'en-tête par l'indication HTML, complétée de la version.

*HTML est une application de la norme SGML.*

*Les caractéristiques des documents HTML sont définies par la DTD HTML, qui précise notamment les jeux de caractères utilisés et les formats disponibles.*

## HTML : versions et extensions

Actuellement, il y a plusieurs versions de DTD HTML.

Les versions 0 et 1 sont désormais obsolètes ; elles prévoyaient principalement les mises en forme de paragraphes et de caractères, l'intégration des images et la définition des ancres permettant la mise en œuvre de liens hypertextes.

La version 2, qui est officiellement en vigueur, comprend de nombreux éléments complémentaires, en particulier pour le positionnement des images ; elle intègre surtout les formulaires, qui permettent à l'utilisateur de réaliser des saisies et de piloter l'exécution de programmes externes sur le serveur. Elle introduit également la notion de niveau de conformance, qui correspond à peu près à la version de DTD utilisée. La valeur 0 du niveau de conformance définit un niveau minimal (formatage simple du texte), la valeur 1 indique quelques extensions, comme les images, alors que la valeur 2 comprend aussi le support des formulaires, permettant à l'utilisateur de saisir des informations qui seront transmises à des programmes externes (bases de données notamment). HTML 3 est actuellement à l'étude et doit incorporer diverses extensions, notamment un mode de description des tableaux et des écritures scientifiques.

Les éditeurs de logiciels clients, comme Netsite pour le client Netscape, anticipent largement sur cette version 3, notamment pour influencer la définition de la DTD correspondante.

Une des difficultés de la définition de la version 3 de la DTD est l'adjonction, notamment par Netsite, de fonctions nouvelles qui sont difficilement intégrables à la norme SGML...

La tendance actuelle des organismes qui publient sur Web est de s'aligner et d'exploiter largement les extensions Netscape (voir par exemple à ce propos la définition du Web du journal *Libération*). Ce comportement peut néanmoins être à l'origine de multiples difficultés dans l'avenir : Netscape devant devenir onéreux, sommes-nous aussi sûrs qu'il restera le client universel, alors que d'autres restent gratuits ? Si la DTD HTML 3 ne reprend pas toutes les extensions Netscape, aurons-nous des services conformes à la norme et des services Netsite ? Est-il sage de favoriser, dans les services mis en place, la future hégémonie d'une société que certains annoncent déjà comme le nouveau Microsoft ? Par ailleurs, Netsite fait des émules, et Microsoft lui-même a annoncé une gamme de produits destinés au Web incorporant nombre d'extensions « propriétaires ».

D'autres acteurs tentent également de définir des extensions nouvelles. Qu'il s'agisse de Java et Hot Java (Sun Microsystems) pour l'exécution de programmes externes, de VRML (Silicon Graphics) pour les animations 3D, de Shockwave pour Director ou d'autres encore, il faut se souvenir que la force du Web c'est sa formidable interopérabilité entre des plates-formes différentes, garantie par la standardisation maintenue par le consortium W3 (<http://www.w3.org>).

---

<sup>1</sup> La DTD HTML est disponible dans le document « Hypertext Markup Language Specification. v. 2.0 », téléchargeable au format Postscript sur le serveur du CERN (voir bibliographie). Un "draft" de la DTD HTML 3 est également disponible.

*Il y a actuellement plusieurs versions de DTD HTML.*

*Les versions 0 et 1, définies par le CERN en 1992-93, sont actuellement obsolètes, et remplacées par la version 2 (1994). Elles comprenaient en particulier les éléments de formatage de texte, les éléments de liens et la gestion des images.*

*La version 2 prévoit notamment la définition de formulaires.*

*Actuellement, la DTD version 3 est en cours de définition. Elle incorpore notamment les tableaux et les formules mathématiques.*

*La standardisation de HTML est maintenue par le consortium W3.*

*Des éditeurs privés, comme Netsite, Sun ou SGI définissent des extensions personnelles à HTML, qui sont destinées principalement à faire pression sur le consortium W3 pour les évolutions à venir.*

## **HTML : texte et balises**

Le langage HTML est fondé sur le principe d'un flux principal textuel, au sein duquel sont placées des balises. Ces balises sont en fait des instructions, au sens informatique du terme, pour l'environnement d'exécution.

Les balises HTML peuvent concerner des indications de mise en forme, mais aussi des instructions de liens vers d'autres documents, au travers d'adresses codées selon le standard URL (*Uniform Resource Locator*).

Les balises HTML doivent être conformes aux spécifications SGML et à la DTD correspondante. Le principe d'une application SGML, définie par sa DTD, est notamment fondé sur le respect d'une syntaxe, en particulier pour ce qui concerne l'inclusion des différents éléments entre eux<sup>2</sup>. Actuellement, de nombreux logiciels de composition de documents HTML n'assurent aucune vérification de ce type ; en cas de syntaxe non respectée, l'exécution par le logiciel client est aléatoire : les commandes non correctement passées seront peut-être ignorées (cas le plus favorable), mais elles peuvent aussi être interprétées de n'importe quelle manière. On comprendra que la simple vérification par l'exécution sur une seule plate-forme et un seul logiciel ne peut suffire.

Pour chacune des commandes décrites, nous avons tenté de préciser les inclusions autorisées, c'est-à-dire les éléments ou commandes qu'elles peuvent contenir. Cependant, ces inclusions sont celles définies par la DTD HTML version 2, et elles ne tiennent donc pas compte des nombreuses extensions actuelles qui relèvent du futur HTML 3.

---

<sup>2</sup> Un élément HTML est défini par une balise (de type <nom\_de\_balise>), suivi d'un texte sur lequel porte la commande désignée par la balise, et d'une balise de fin d'élément (du type </nom\_de\_balise>). Le texte inclus entre les balises de début et de fin peut également comporter d'autres éléments, mais cela suppose le respect d'une syntaxe précise au niveau des inclusions.

*Le principe de HTML est fondé sur un flux textuel dans lequel sont incorporées des balises de formatage et de définition d'hyperliens.*

*Les principes de SGML imposent de considérer les règles d'écriture de ces balises comme une syntaxe, dont le respect est la garantie d'une bonne exécution, sur tous types de configurations.*

## Les URL

L'information contenue dans un lien doit indiquer de manière non ambiguë où et comment atteindre la ressource référencée. Pour cela, W3 utilise une structure d'adressage nommée URL (*Uniform Resource Locator*), qui est l'extension au niveau de l'Internet de la notion de nom de fichier sur une machine.

Un URL permet d'adresser de façon précise toute ressource accessible sur l'Internet.

Un URL comprend toutes les informations nécessaires à la localisation et à l'exploitation d'une ressource du réseau, notamment le protocole à suivre, le nom de la machine, le chemin et le nom du fichier à utiliser, mais aussi toutes informations complémentaires nécessaires pour le protocole concerné, comme le nom d'utilisateur et le mot de passe pour un transfert de fichiers, ou encore l'adresse E-mail pour l'envoi d'un message.

De plus il existe deux sortes d'URL :

- Les URL absolus ;
- Les URL relatifs.

Les URL absolus répondent à la syntaxe suivante :

méthode://nom\_machine[:port]/chemin/nom\_fichier[#ancree|?liste\_de\_paramètres]

Le champ **méthode** indique le protocole utilisé. Actuellement, il en existe de nombreux, notamment: file, ftp, http, telnet, gopher, wais, news, mailto (liste non exhaustive).

Le champ **port** permet d'indiquer la porte<sup>3</sup> à utiliser sur le serveur pour ouvrir un dialogue avec le daemon (logiciel serveur). Ce champ est facultatif, dès lors que l'on respecte les ports standards (par exemple port 80 pour le daemon http).

Le **chemin** d'accès au fichier est donné à partir de la racine logique attribuée au daemon sollicité (le champ « **méthode** » définit le daemon appelé ; chaque daemon définit, dans son fichier de paramétrage, une adresse de base pour ses fichiers de données, qui constitue la racine logique).

**#ancree** est optionnel. Il permet d'indiquer un signet dans le document cible. Le document sera alors ouvert directement à la position du signet (par défaut, l'ouverture se fait au début du document).

**?Liste\_de\_paramètres** est optionnel. Il permet, lorsque le fichier appelé est un programme exécutable, de passer des paramètres sur la ligne de commande transmise par le daemon.

Les URL relatifs consistent en un nom de fichier, éventuellement complété de son chemin absolu ou relatif ([/chemin]/[fichier]). Un chemin absolu est défini à partir du répertoire racine du daemon appelé (en fonction de « **méthode** »), qui peut être différent du répertoire ROOT de la machine ; un chemin relatif est défini à partir du répertoire courant du fichier utilisé (les raccourcis « . » et « .. » indiquant le répertoire courant et le répertoire parent sont valides). On utilise généralement un URL relatif au sein d'un document pour référencer un autre document localisé sur le même serveur et accessible par le même protocole.

La définition d'une méthode unifiée de nommage permet de simplifier l'écriture de documents HTML.

*Les URL définissent des règles de description de l'adresse d'une ressource sur Internet.*

---

<sup>3</sup> La « porte » d'un programme sur une machine est une notion issue du système Unix (*system ports*) Il s'agit en fait d'une plage d'adresses spécifique désignée pour les communications d'entrées/sorties du programme. Le protocole TCP, utilisé par Internet (TCP/IP) utilise le dialogue par portes.