World Wide Web

Le ***World Wide Web*** (WWW), littéralement la « **toile (d’araignée) mondiale** », communément appelé le **Web**, et parfois la **Toile**, est un système hypertexte public fonctionnant sur Internet. Le Web permet de consulter, avec un navigateur, des pages accessibles sur des sites. L’image de la toile d'araignée vient des hyperliens qui lient les pages web entre ellesnote 1.

Le Web n’est qu'une des applications d’Internet ; distincte d’autres applications comme le courrier électronique, la messagerie instantanée, et le partage de fichiers en pair à pair. Le Web a été inventé par Tim Berners-Lee plusieurs années après Internet, mais c’est lui qui a rendu les médias grand public attentifs à Internet. Depuis, le Web est fréquemment confondu avec Internet1 ; en particulier, le mot *Toile* est souvent utilisé dans les textes non techniques sans qu’il désigne clairement le Web ou Internet.



Mondial par définition

**Terminologie**

**Termes désignant le *World Wide Web***

Le *World Wide Web* est désigné par de nombreux noms et abréviations synonymes : *WorldWideWeb*, *World Wide Web*, *World-wide Web*, *Web*, *WWW*, *W3*, *Toile d’araignée mondiale*, *Toile mondiale*, *Toile*. Certains ont disparu.

Le nom du projet originel était *WorldWideWeb*2. Les mots ont été rapidement séparés en *World Wide Web* pour améliorer la lisibilité. Le nom *World-Wide Web* a également été utilisé par les inventeurs du Web, mais le nom désormais préconisé par le *World Wide Web Consortium (W3C)* sépare les trois mots sans trait d’union3. Bien que « mondial » s’écrive *world-wide* ou *worldwide* en anglais, l’orthographe *World Wide Web* et l’abréviation *Web* sont maintenant bien établies.

En inventant le Web, Tim Berners-Lee avait aussi pensé à d’autres noms, comme *Information Mesh* (maillage d’informations), *Mine of Information* ou encore *The Information Mine* (la mine d’informations, dont le sigle serait *Tim*).

Le sigle *WWW* a été largement utilisé pour abréger *World Wide Web* avant que l’abréviation *Web* prenne le pas. *WWW* se prononce souvent *trois double V*, *triple double V*, *vévévé* ou *wéwéwé* (en Belgique). La prononciation laborieuse (en français comme en anglais) de *WWW* a sans doute précipité son déclin à l'oral. À l'écrit, les lettres *www* restent très utilisées dans les adresses Web et quelques autres conventions techniques. *WWW* est parfois abrégé en *W3*, abréviation qu’on retrouve dans le sigle *W3C* du *World Wide Web Consortium*. Dans la seconde moitié des années 1990, une blague répandue prétendait que *WWW* signifiait *World Wide Wait*, soit « attente mondiale », car le réseau Internet était engorgé par la popularité grandissante du Web.

Pour écrire « le web », l’usage de la minuscule est de plus en plus courant. L’Office québécois de la langue française préconise la majuscule4, le Journal officiel français préconise « la toile d’araignée mondiale5 ». Cet article fait la distinction entre « le Web » et « un *web* », aussi la majuscule est toujours utilisée pour désigner le Web.

**Termes rattachés au web**

La terminologie propre au web contient plusieurs dizaines de termes. Ce chapitre expose ceux qui sont utilisés dans cet article.

L’expression **en ligne** signifie « connecté à un réseau », en l’occurrence le réseau informatique Internet. Cette expression n’est pas propre au web, on la retrouve à propos du téléphone.

Un **hôte** est un ordinateur en ligne. Chaque hôte d’Internet est identifié par une adresse IP à laquelle correspondent zéro, un ou plusieurs **noms d’hôte**. Cette terminologie n’est pas propre au web, mais à Internet.

Une **ressource du World Wide Web** est une entité informatique (texte, image, forum Usenet, boîte aux lettres électronique, etc.) **accessible** indépendamment d’autres ressources. Une ressource en accès **public** est librement accessible depuis Internet. Une ressource **locale** est présente sur l’ordinateur utilisé, par opposition à une ressource **distante** (ou en ligne), accessible à travers un réseau.

On ne peut accéder à une ressource distante qu’en respectant un **protocole de communication**. Les fonctionnalités de chaque protocole varient : réception, envoi, voire échange continu d’informations.

**HTTP** (pour *HyperText Transfer Protocol*) est le protocole de communication communément utilisé pour transférer les ressources du web. **HTTPS** est la variante sécurisée de ce protocole.

Une **URL** (pour *Uniform Resource Locator*) **pointe** sur une ressource. C’est une chaîne de caractères permettant d’indiquer un protocole de communication et un emplacement pour toute ressource du web.

Un **hyperlien** (ou **lien**) est un élément dans une ressource associé à une URL. Les hyperliens du web sont orientés : ils permettent d’aller d’une source à une destination. Seule la ressource à la source contient les données définissant l’hyperlien, la ressource de destination n’en porte aucune trace. Il existe deux types d’hyperlien : ceux du premier type doivent être activés pour accéder à la destination ; ceux du second causent un accès automatique à la destination.

**HTML** (pour *HyperText Markup Language*) et XHTML (*Extensible HyperText Markup Language* ) sont les langages informatiques permettant de décrire le contenu d’un document (titres, paragraphes, disposition des images, etc.) et d’y inclure des hyperliens. Un **document HTML** est un document décrit avec le langage HTML. Les documents HTML sont les ressources les plus consultées du web.

Dans un mode de communication **client-serveur**, un **serveur** est un hôte sur lequel fonctionne un *logiciel serveur* auquel peuvent se connecter des *logiciels clients* fonctionnant sur des *hôtes clients*.

Un **serveur web** est un hôte sur lequel fonctionne un **serveur HTTP** (ou **serveur web**). Un serveur web **héberge** les ressources qu’il dessert.

Un **navigateur web** est un logiciel client HTTP conçu pour accéder aux ressources du web. Sa fonction de base est de permettre la consultation des documents HTML disponibles sur les serveurs HTTP. Le support d’autres types de ressource et d’autres protocoles de communication dépend du navigateur considéré.

Une **page web** (ou **page**) est un document destiné à être consulté avec un navigateur web. Une page web est toujours constituée d’une ressource centrale (généralement un document HTML) et d’éventuelles ressources liées automatiquement accessibles (par exemple, des images).

Un **éditeur HTML** (ou **éditeur web**) est un logiciel conçu pour faciliter l’écriture de documents HTML et de pages web en général.

Un **site web** (ou **site**) est un ensemble de pages web et d’éventuelles autres ressources, liées dans une structure cohérente, publiées par un propriétaire (une entreprise, une administration, une association, un particulier, etc.) et hébergées sur un ou plusieurs serveurs web.

**Visiter** un site web signifie « consulter ses pages ». Le terme **visite** vient du fait que l’on consulte généralement plusieurs pages d’un site, comme on visite les pièces d’un bâtiment. La visite est menée par un **utilisateur** (ou **visiteur** ou **internaute**). La mesure d’audience est obtenue en copiant le code en javascript d’un lien vers le site d’un prestataire spécialisé suivant la technique du marqueur à distance.

Une **adresse web** est une URL de page web, généralement écrite sous une forme simplifiée limitée à un nom d’hôte. Une adresse de site web est en fait l’adresse d’une page du site prévue pour accueillir les visiteurs.

Un **hébergeur web** est une entreprise de services informatiques hébergeant (mettant en ligne) sur ses serveurs web les ressources constituant les sites web de ses clients.

Une **agence web** est une entreprise de services informatiques réalisant des sites web pour ses clients.

L’expression **surfer sur le web** signifie « consulter le web ». Elle a été inventée pour mettre l’accent sur le fait que consulter le web consiste à suivre de nombreux hyperliens de page en page. Elle est principalement utilisée par les médias ; elle n’appartient pas au vocabulaire technique.

Un **annuaire web** est un site web répertoriant des sites web.

Un **portail web** est un site web tentant de regrouper la plus large palette d’informations et de services possibles dans un site web. Certains portails sont thématiques.

Un **agrégateur web** est un site web qui sélectionne, organise et, éventuellement, valide des pages (URL) concernant un sujet précis et de les met en forme de façon attractive.

Un **service web** est une technologie client-serveur fondée sur les protocoles du web.

**Histoire**

[](http://commons.wikimedia.org/wiki/File:Cailliau_Abramatic_Berners-Lee_10_years_WWW_consortium.png)

Robert Cailliau, Jean-François Abramatic et Tim Berners-Lee lors du 10e anniversaire du *WWW Consortium*.

Tim Berners-Lee (TBL) travaille comme informaticien dans le bâtiment 31 (46.2325, 6.04506) de l’Organisation européenne pour la recherche nucléaire (CERN), situé à la frontière franco-suisse, lorsqu’il propose, en 1989, de créer un système hypertexte distribué sur le réseau informatique pour que les collaborateurs puissent partager les informations au sein du CERN7. Cette même année, les responsables du réseau du CERN décident d’utiliser le protocole de communication TCP/IP et le CERN ouvre sa première connexion extérieure avec Internet8.

L’année suivante, l’ingénieur système Robert Cailliau se joint au projet d’hypertexte au CERN, immédiatement convaincu de son intérêt, et se consacre énergiquement à sa promotion9. Tim Berners-Lee et Robert Cailliau sont reconnus comme les deux personnes à l’origine du *World Wide Web*. Tim Berners-Lee crée le premier navigateur et éditeur web, appelé WorldWideWeb et le premier serveur HTTP appelé CERN httpd.

Jusqu’en 1993, le web est essentiellement développé sous l’impulsion de Tim Berners-Lee et Robert Cailliau. Les choses changent avec l’apparition de NCSA Mosaic, un navigateur web développé par Eric Bina et Marc Andreessen au *National Center for Supercomputing Applications* (NCSA), dans l’Illinois. NCSA Mosaic jette les bases de l’interface graphique des navigateurs modernes en intégrant les images au texte et cause un accroissement exponentiel de la popularité du web. Certains développeurs de Mosaic, crééront ensuite Netscape Navigator qui apporte le parallélisme, permettant également d'utiliser les pages avec image sur des connexions bas débit. Celui ci d'évolutions en évolutions est devenu aujourd'hui Mozilla Firefox. Le NCSA produit également le NCSA HTTPd, un serveur HTTP qui évoluera en Apache HTTP Server, le serveur HTTP le plus utilisé depuis 1996.



Le logo de Mozilla Firefox

Le 30 avril 1993, le CERN verse dans le domaine public toutes les technologies développées autour du www10. Pour l'occasion, l'organisation place en ligne ce qui peut être considéré comme étant le premier site en "www"11, hébergé sur la station NeXT de TBL10.

En 1994, Netscape Communications Corporation est fondée avec une bonne partie de l’équipe de développement de NCSA Mosaic. Sorti fin 1994, Netscape Navigator supplante NCSA Mosaic en quelques mois.



Netscape Navigator

En 1995, Microsoft essaie de concurrencer Internet avec The Microsoft Network (MSN) et échoue. Fin 1995, après la sortie de Windows 95 sans le moindre navigateur web préinstallé, Microsoft lance avec Internet Explorer la guerre des navigateurs contre Netscape Navigator.

### Évolution de l'usage

En créant le logiciel *WorldWideWeb*, Tim Berners-Lee a créé à la fois le premier navigateur web et le premier éditeur web, car il voulait faire du web un média collaboratif, dans lequel tous les acteurs consultent et créent l'information. Cependant, le web s'est immédiatement orienté en un média de diffusion d'information global plutôt que de collaboration22.

Dans la première moitié des années 1990, le concept de site web à la racine d'un nom de domaine stable n'était pas établi, à commencer par la première page web dont l'adresse était http://nxoc01.cern.ch/hypertext/WWW/TheProject.html. Les sites étaient souvent mis en place dans des départements techniques par des employés et des étudiants, et les URL changeaient au gré des changements de personnes et d'infrastructure. En outre, il n'existait pas de moteur de recherche efficace. Aussi de nombreuses pages étaient des listes de liens sur les pages préférées de l'auteur de la page. Cette propriété du web sera d'ailleurs exploitée plus tard par les créateurs de Google pour calculer la pertinence des pages, puis détournée par les fermes de liens. Mais bien avant cela, en janvier 1994, Yahoo! est créé et devient rapidement le plus grand annuaire web. Comme les créateurs de Yahoo! étaient des étudiants de l'université Stanford, L'URL originale était http://akebono.stanford.edu/yahoo, et ce n'est qu'en janvier 1995 que le domaine yahoo.com est créé.



Logo 2013 de Yahoo.com

Dans la seconde moitié des années 1990, le web devient populaire, et toutes les grandes entreprises, organisations, écoles, administrations, ouvrent un site web. Les moteurs de recherche deviennent efficaces, notamment avec l'apparition d'Altavistanote 2 en décembre 1995, et pour finir Google en 1998. Dans cette phase de développement du média, un flot d'information *top-down* prédomine : un site web est fait pour diffuser les informations de son propriétaire. Les interactions s'arrêtent souvent à la recherche et au commerce en ligne. Bien sûr, le courrier électronique, la messagerie instantanée, et les forums de discussion existaient depuis plus longtemps que le web, mais ils fonctionnaient principalement avec des protocoles et logiciels spécifiques (voir Catégorie:Logiciel de messagerie instantanée, Catégorie:Logiciel de courrier électronique, Catégorie:Client Usenet).

#### Web 2.0

Avec les années 2000 les notions de blog, de wiki (en 2001, lancement de la Wikipédia en anglais) et de réseautage social (Myspace en 2003, Facebook en 2004) deviennent populaires. Le contenu généré par les utilisateurs se répand (Wikipédia, YouTube en 2005, Twitter en 2006). La technologie Ajax (1998, théorisée en 2005) commence à être largement utilisée pour créer des applications complètes qui tiennent dans une seule page web (Google Maps en 2004). L'expression Web 2.0, largement popularisée au milieu des années 2000, désigne cette transition dans le flux de l'information et la manière d'utiliser le web. Le succès de l'expression Web 2.0 a conduit de nombreuses personnes à appeler Web 2.5, 3.0, 4.0, etc. leur vision du Web de l'avenir.

#### Web 3.0

L'expression Web 3.0 est utilisée en futurologie à court terme pour désigner l'internet qui suit le Web 2.0 et constitue l'étape à venir du développement du World Wide Web. Son contenu réel n'est pas défini de manière consensuelle, chacun l'utilisant pour désigner sa propre vision du futur d'internet.

## Architecture

### Modèle mathématique

Le *World Wide Web*, en tant qu’ensemble de ressources hypertextes, est modélisable en *graphe orienté possédant des cycles* avec les ressources pour *sommets* et les hyperliens pour *arcs*. Comme le graphe est orienté, certaines ressources peuvent constituer des *puits*, ou moins formellement des cul-de-sac23 : il n’existe aucun *chemin* vers le reste du web. À l’inverse, certaines ressources peuvent constituer des *sources* : il n’existe aucun *chemin* depuis le reste du web.

Les analyses ont montré que la structure du web répondait au modèle des réseaux libres d'échelle24 présent dans la plupart des réseaux sociaux. Cela se traduit par la présence de moyeux, les hubs, vers lesquels convergent les liens hypertextes : ce sont les sites les plus importants qui constituent le squelette du web.

Techniquement, rien ne distingue le *World Wide Web* d’un quelconque autre web utilisant les mêmes technologies. Ainsi, d’innombrables webs privés existent. Dans la pratique, on considère qu’une page d’un site web populaire, comme un annuaire web, fait partie du web. Le web peut alors être défini comme étant l’ensemble des ressources et des hyperliens que l’on peut récursivement découvrir à partir de cette page, ce qui exclut les sources et les webs privés.

### Exploration du web et web profond

L’exploration récursive du web à partir de ressources bien choisies est la méthode de base programmée dans les robots d’indexation des moteurs de recherche. En 2004, les moteurs de recherche indexent environ 4 milliards de ressources.

Le web profond, ou web invisible, est la partie du web qui n’est pas indexée et donc introuvable avec les moteurs de recherche généralistes. Une étude publiée en 2001 indiquait que la partie invisible du web représente plus de 99 % du web25. Le web profond comprend notamment les ressources suivantes :

* les ressources inaccessibles au public, donc aux robots, notamment les pages administratives ou payantes, protégées par un mot de passe ;
* les ressources qui ne sont pas communiquées par des protocoles de communication pris en charge par les robots (souvent ils ne prennent en charge que HTTP et HTTPS) ;
* les ressources dont le format de données n'est pas pris en charge par le robot ;
* les ressources listées dans un fichier d’exclusion des robots ;
* les ressources exclues par le robot car elles sont conçues pour abuser du référencement (spamdexing) ;
* les ressources exclues par le robot car elles sont considérées comme trop peu pertinentes (par exemple si un site contient des millions de ressources qui ne sont liées par aucun autre site) ;
* les ressources vers lesquelles les hyperliens sont créés dynamiquement en réponse aux interrogations des visiteurs.

Ces dernières ressources proviennent généralement de bases de données et constituent la partie la plus importante du web profond.

### Serveurs publics

L’exploration récursive n’est pas le seul moyen utilisé pour indexer le web et mesurer sa taille. L’autre solution consiste à mesurer l’infrastructure informatique connectée à Internet pour héberger des sites web. Au lieu de suivre des hyperliens, cette méthode consiste à utiliser les noms de domaine enregistrés dans le *Domain Name System* et essayer de se connecter à tous les serveurs web potentiels. C’est notamment la méthode utilisée par la société Netcraft, qui publie régulièrement les résultats de ses explorations, dont les mesures de popularité des serveurs HTTP. Cette mesure porte plus sur l’utilisation des technologies du web que sur le web lui-même. Elle permet notamment de trouver des sites publics qui ne sont pas liés au *World Wide Web*.

### Intranets et *webs* privés

Un site web mis en ligne sur un intranet est privé, car le public ne peut pas accéder à un intranet.

En outre, si l'on met en ligne un site web sur Internet en omettant de créer des liens depuis au moins une page existante du *World Wide Web*, alors ce site constitue un web isolé. Il est virtuellement privé, car le public ne peut pas le découvrir en suivant des hyperliens.

### Archivage

Le web change constamment : les ressources ne cessent d’être créées, modifiées et supprimées. Il existe quelques initiatives d’archives du web dont le but est de permettre de retrouver ce que contenait un site à une date donnée. Le projet Internet Archive est l’un d’eux.

## Types de ressource

Les divers types de ressource du web ont des usages assez distincts :

* les ressources constituant les pages web : documents HTML, images JPEG ou PNG ou GIF, scripts JavaScript, feuilles de style CSS, sons, animations, vidéo ;
* les ressources accessibles depuis une page web mais consultables avec une interface particulière : applet ;
* les ressources conçues pour être consultées séparément : documents (PDF, PostScript, Word, etc.), fichier texte, images de tout type, morceaux de musique, vidéo, fichiers à sauvegarder ;
* les ressources appartenant à des systèmes qui ont une existence indépendante du web, mais vers lesquelles il est possible de créer un hyperlien : forums Usenet, boîtes aux lettres électronique, fichiers locaux.

### Documents HTML

[](http://commons.wikimedia.org/wiki/File:Page_Web.png)

Une page web

Le document HTML est la principale ressource d’une page web, celle qui contient les hyperliens, qui contient et structure le texte, qui lie et dispose les ressources multimédias. Un document HTML contient uniquement du texte : le texte consulté, le texte en langage HTML plus d’éventuels autres langages de script ou de style.

La présentation de documents HTML est la principale fonctionnalité d’un navigateur web. HTML laisse au navigateur le soin d’exploiter au mieux les capacités de l’ordinateur pour présenter les ressources. Typiquement, la police de caractère, la longueur des lignes de texte, les couleurs, etc, doivent être adaptées au périphérique de sortie (écran, imprimante, etc).

### Multimédia

Les éléments multimédias proviennent toujours de ressources indépendantes du document HTML. Les documents HTML contiennent des hyperliens pointant sur les ressources multimédias, qui peuvent donc être éparpillées sur Internet. Les éléments multimédias liés sont automatiquement transférés pour présenter une page web.

Seul l’usage des images et des petites animations est standardisé. Le support du son, de la vidéo, d’espaces tridimensionnels ou d’autres éléments multimédias repose encore sur des technologies non standardisées. De nombreux navigateurs web proposent la possibilité de greffer des logiciels (plugin) pour étendre leurs fonctionnalités, notamment le support de types de média non standard.

Les flux (audio, vidéo) nécessitent un protocole de communication au fonctionnement différent de HTTP. C’est une des raisons pour lesquelles ce type de ressource nécessite souvent un plugin et est mal intégré aux pages web.

### Images

Ce chapitre concerne les images intégrées aux pages web.

L’usage du format de données JPEG est indiqué pour les images naturelles, principalement les photographies.

L’usage du format de données PNG est indiqué pour les images synthétiques (logos, éléments graphiques). Il est aussi indiqué pour les images naturelles, mais uniquement lorsque la qualité prime totalement sur la durée du transfert.

L’usage du format de données GIF est indiqué pour les petites animations. Pour les images synthétiques, la popularité ancienne de GIF le fait souvent préférer à PNG. Cependant, GIF souffre de quelques désavantages, notamment la limitation du nombre de couleurs et un degré de compression généralement moindre. En outre une controverse a entouré l’usage de GIF de 1994 à 2004 car Unisys a fait valoir un brevet couvrant la méthode de compression.

L’usage d’images de format de données XBM est obsolète.

### Vidéo

Jusque dans les années 2000, la consultation de musique et vidéo demandait l'installation d'un programme ad hoc (un plugin) pour étendre les fonctionnalités du navigateur web. La très grande diffusion du plugin Flash Player a finalement rendu la consultation vidéo aussi simple que celle des images. Finalement, la cinquième version langage HTML (HTML 5) a intégré la vidéo.

### Scripts et animations

Un langage de script permet d’écrire le texte d’un programme directement exécuté par un logiciel. Dans le cadre du web, un script est exécuté par un navigateur web et programme des actions répondant à l’usage que le visiteur fait de la page web consultée. Un script peut être intégré au document HTML ou provenir d’une ressource liée. Le premier langage de script du web fut JavaScript, développé par Netscape. Ensuite Microsoft a développé une variante concurrente sous le nom de JScript. Finalement, la norme ECMAScript a été proposée pour la syntaxe du langage, et les normes DOM pour l’interface avec les documents.

De technologie aux capacités d'actions très limitées à ses débuts, le langage JavaScript est devenu capable d'exécuter toutes les applications imaginables : traitement de texte, jeu vidéo, émulateur, etc.

Encore plus que pour la vidéo, le plugin Flash Player est devenu très largement utilisé pour la diffusion d'animations. Parfois, des sites entiers sont réalisés en Flash.

### Styles

Le langage CSS a été développé pour gérer en détail la présentation des documents HTML. Le texte en langage CSS peut être intégré au document HTML ou provenir de ressources liées, les feuilles de style. Cette séparation permet une gestion séparée de l’information (contenue dans des documents HTML) et de sa présentation (contenue dans des feuilles de style). On parle aussi de « séparation du fond et de la forme ».

### Autres

La gestion des autres types de ressource dépend des logiciels installés sur l’hôte client et de leurs réglages.

Lorsque le logiciel correspondant est disponible, les documents et images de tout type sont généralement automatiquement présentés, selon des modalités (fenêtrage, dialogues) dépendant du navigateur web et du logiciel gérant le type. Lorsque le type de la ressource n’est pas géré, il est généralement possible de la sauver dans un fichier local.

Pour gérer les ressources de systèmes différents du web comme le courrier électronique, les navigateurs font habituellement appel à des logiciels séparés. Si aucun logiciel ne gère un type de ressource, un simple message d’erreur l’indique.

## Conception

### Universalité

Le web a été conçu pour être accessible avec les équipements informatiques les plus divers : station de travail, terminal informatique en mode texte, ordinateur personnel, téléphone portable, etc. Cette universalité d’accès dépend en premier lieu de l’universalité des protocoles Internet. En second lieu, elle dépend de la flexibilité de présentation des pages web, offerte par HTML. En outre, HTTP offre aux navigateurs la possibilité de négocier le type de chaque ressource. Enfin, CSS permet de proposer différentes présentations, sélectionnées pour leur adéquation avec l’équipement utilisé.

Le W3C a pour cela créé des normes dans le but de permettre l'indépendance des outils qui servent à créer du contenu avec ceux qui servent à le lire. On appelle cela l'interopérabilité.

L’accessibilité du web pour les individus handicapés est aussi l’objet d’attentions particulières comme la *Web Accessibility Initiative*.

### Décentralisation

Les technologies du web n’imposent pas d’organisation entre les pages web, ni à fortiori entre les sites web. Toute page du web peut contenir des hyperliens vers toute autre ressource accessible d’Internet. L’établissement d’un hyperlien ne requiert absolument aucune action du côté de la ressource pointée. Il n’y a pas de registre centralisé d’hyperliens, de pages ou de sites. Le seul registre utilisé est celui du DNS ; c’est une base de donnée distribuée qui répertorie les hôtes, permet de traduire en adresse IP le nom de domaine contenu dans certains hyperliens et qui est utilisée par tous les systèmes accédant à Internet.

Cette conception décentralisée devait favoriser, et a favorisé, une augmentation rapide de la taille du web. Elle a aussi favorisé l’essor de sites spécialisés dans les informations sur les autres sites : les annuaires et les moteurs de recherche. Sans ces sites, la recherche d’information dans le web serait extrêmement laborieuse. La démarche inverse, le portail web, tente de concentrer un maximum d’informations et de services dans un seul site.

Une faiblesse de la décentralisation est le manque de suivi lorsqu’une ressource est déplacée ou supprimée : les hyperliens qui la pointaient se retrouvent *cassés*. Et cela n’est visible qu’en suivant l’hyperlien, le résultat le plus courant étant le message d’erreur 404.

## Technologies

### Pré-existantes

Le web repose sur les technologies d’Internet, notamment TCP/IP pour assurer le transfert des données, DNS pour convertir les noms d’hôte en adresses IP et MIME pour indiquer le type des données. Les standards de codage de caractères et les formats d’image numérique GIF et JPEG ont été développés indépendamment.

### Spécifiques

Trois technologies ont dû être développées pour le *World Wide Web* :

* les URL pour pouvoir identifier toute ressource dans un hyperlien ;
* le langage HTML pour écrire des pages web contenant des hyperliens ;
* le protocole de communication HTTP utilisé entre les navigateurs et les serveurs web, qui permet d’indiquer le type MIME des ressources transférées.

Ces premières technologies ont été normalisées comme les autres technologies d’Internet : en utilisant le processus des *Request for Comments*. Cela a donné le RFC 1738 pour les URL, le RFC 1866 pour HTML 2.0 et le RFC 1945 pour HTTP/1.0.

Le *World Wide Web Consortium* (W3C) a été fondé en 1994 pour développer et promouvoir les nouveaux standards du web. Son rôle est notamment de veiller à l’universalité des nouvelles technologies. Des technologies ont également été développées par des entreprises privées.

### Actuelles

Les principaux standards actuels sont :

* XML 1.0 développé pour donner aux langages de balises, dont HTML, une syntaxe plus simple que SGML ;
* HTML 4.01 fondé sur SGML, et XHTML 1.0 fondés sur XML ;
* le RFC 2396 (*Uniform Resource Identifiers*), qui recouvre les URL ;
* le RFC 2616 (HTTP/1.1) ;
* les feuilles de styles en cascade CSS level 1 et level 2 ;
* les modèles de document DOM level 1 et level 2 ;
* le langage de script JavaScript pour manipuler les documents ;
* les formats d’image numérique PNG, JPEG et GIF.

### Technologies serveur

Outre les protocoles de communication et formats de données échangés sur le Web, plusieurs techniques propres au Web sont mises en œuvre pour faire fonctionner les serveurs HTTP (ou serveurs web). Comme ces techniques ne concernent pas les logiciels client du Web, elles ne sont pas standardisée par le World Wide Web Consortium.

#### Serveur HTTP

Le serveur HTTP est chargé d'assurer la communication entre le poste client et les ressources des sites. Il doit être capable d'aiguiller, sur le serveur, en fonction de la requête HTTP, reçue de la part du client, vers les bonnes ressources. Il peut s'agir de ressources statiques, situées sur le système de fichier du serveur ou un système de fichier déporté, comme par exemple NFS, ou des contenus dynamiques par différents biais. La fourniture des pages dynamiques est déléguée à des applications autonomes sur le serveur. Il doit également être capable de gérer les erreurs, si les ressources ne sont pas trouvées ou s'il y a une erreur dans la production de la ressource, en retournant le message adéquat au client.



Les serveurs Web

Le premier serveurs web fut CERN httpd crée par l'inventeur du World Wide Web. Il est rapidement tombé en désuétude.

Aujourd'hui, les principaux serveurs Web utilisés sont Apache HTTP Server, serveur dominant du marché depuis une dizaine d'année, Nginx utilisé sur les sites à plus forte audience, Microsoft IIS et Google Web Server.

#### Communication entre le serveur HTTP et les applications serveur autonomes

* Le standard *Common Gateway Interface* (CGI) est un protocole de communication inter-processus entre le serveur HTTP et des applications externes situées également du côté serveur. Ce standard est géré par l'IETF.
* Le standard FastCGI remplace aujourd'hui majoritairement le CGI. Il permet de séparer plus efficacement le serveur HTTP des applications et de mieux en contrôler le nombre d'instances des applications dynamiques du côté serveur.

Certains langages couramment utilisés avec les service HTTP, comme PHP, Java et Ruby, utilisent également des technologies particulières pour la communication avec le serveur.

* Pour PHP, mod\_php pour Apache a souvent été utilisé à la place de CGI ; il tend aujourd'hui à être remplacé par le protocole fastCGI
* En Java, le projet Apache Jakarta a permis de réunir plusieurs outils. Dans cette architecture, le module Apache mod\_jk permet de faire la liaison avec le serveur d'application Tomcat qui exécute les Java Servlet et JavaServer Pages.
* En Ruby, Phusion Passenger, fonctionnant avec Apache et Nginx est un des principaux outils de communication entre le service HTTP et les applications.

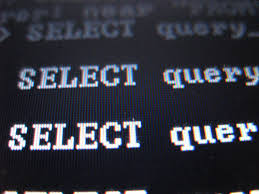
#### Langages d'application dynamiques côté serveurs

En pratique, CGI et fastCGI permettent d'utiliser n'importe quel langage, cependant, on peut citer certains langages de script qui sont devenus les plus populaires, utilisant du CGI ou une autre technique pour communiquer avec le serveur.

* Le langage de programmation PHP *(PHP: Hypertext Preprocessor)* a été développé pour générer les pages web. Il jouit d’une forte intégration avec le serveur HTTP et les langages HTML. C'est aujourd'hui le langage le plus utilisé côté serveur avec presque 70 % des sites en 201026.
* Le moteur ASP *(Active Server Pages)* a été développé par Microsoft pour interpréter du langage de script dans le serveur IIS *(Internet Information Services)*, il est le second langage le plus utilisé avec un peu moins de 30 % des sites en 201026.
* Le langage Java, dont les pages sont servies par Apache Tomcat est ensuite très utilisé, particulièrement dans le domaine bancaire, la tendance est à y inclure du langage Ruby via le module jRuby. Java était utilisé sur un peu moins de 1 % des serveurs en 201026.
* Le langage Ruby, avec principalement la serveur d'application Ruby on Rails avec 0,5 % en 201026.
* Devant la popularité grandissante de JavaScript coté client, liée à Ajax puis HTML5, le côté serveur a suivi avec la plateforme logicielle Node.js, dédiée aux applications serveur.
* Au début des CGI, Perl était très utilisé côté serveur, en raison de ses affinités avec les administrateurs système et réseau, et de sa dominance dans ce domaine.

#### Bases de données

Les bases de données sont également une partie importante de la génération des sites à contenu dynamique.



Langage de requêtes

Les principaux types de base de données sont :

* De type SQL ; MySQL est la plus populaire, suivie de Microsoft SQL Server, PostgreSQL, SQLite, Berkeley DB et Oracle.
* De type NoSQL ; MongoDB est la plus utilisée, mais on peut également citer CouchDB, Amazon SimpleDB ou BigTable utilisé par Google.