

# World Wide Web ( WWW )

---

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 elles.

Le Web n'est qu'une des applications d'Internet<sup>1</sup> ; 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 et Robert Cailliau 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 Internet<sup>2</sup> ; en particulier, le mot *Toile* est souvent utilisé dans les textes non techniques sans qu'il désigne clairement le Web ou Internet.

## Sommaire

### 1 Terminologie

- 1.1 Termes généraux
- 1.2 Termes rattachés au web
  - 1.2.1 Serveur HTTP
  - 1.2.2 Client HTTP
  - 1.2.3 Documents
  - 1.2.4 Consultation et mesure
  - 1.2.5 Métiers du web
- 1.3 Types de services web

### • 2 Histoire

- 2.1 Chronologie
- 2.2 Évolution de l'usage
  - 2.2.1 Web 2.0
  - 2.2.2 HTML5
  - 2.2.3 Web 3.0
  - 2.2.4 Répartition du trafic web

### • 3 Architecture

- 3.1 Modèle mathématique
- 3.2 Exploration du web et web profond
- 3.3 Serveurs publics
- 3.4 Intranets et *webs* privés
- 3.5 Archivage

- 4 Types de ressource
    - 4.1 Documents HTML
    - 4.2 Multimédia
    - 4.3 Images
    - 4.4 Vidéo
    - 4.5 Scripts et animations
    - 4.6 Styles
    - 4.7 Autres
  - 5 Conception
    - 5.1 Universalité
    - 5.2 Décentralisation
  - 6 Technologies
    - 6.1 Pré-existantes
    - 6.2 Spécifiques
    - 6.3 Actuelles
    - 6.4 Technologies serveur
      - 6.4.1 Serveur HTTP
      - 6.4.2 Langages d'application dynamiques côté serveurs
      - 6.4.3 Bases de données
  - 7 Notes et références
    - 7.1 Notes
    - 7.2 Références
  - 8 Voir aussi
    - 8.1 Articles connexes
      - 8.1.1 Concepts généraux
      - 8.1.2 Concepts du web
      - 8.1.3 Technologies
      - 8.1.4 Logiciels serveurs
      - 8.1.5 Logiciels clients
      - 8.1.6 Acteurs
      - 8.1.7 Applications
    - 8.2 Liens externes
-

# Terminologie

---

## Termes généraux

---

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*<sup>3</sup>. 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'union<sup>4</sup>. 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 abrégé *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 majuscule<sup>5</sup>, le Journal officiel français préconise « la toile d'araignée mondiale<sup>6</sup> ».

---

## Termes rattachés au web

---

L'expression « en ligne » signifie « connecté à un réseau », en l'occurrence le réseau informatique Internet. Cette expression n'est donc pas propre au web, mais à Internet dans sa globalité, on la retrouve également à propos des réseaux téléphoniques.

Un hôte est un ordinateur connecté à un réseau.

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 tout Internet. Lorsqu'une ressource est présente sur l'hôte de l'utilisateur, elle est dite locale ; par opposition elle est distante si elle est présente sur un hôte différent.

Lorsque celle-ci est connectée et disponible au réseau, elle est en ligne. 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 avec authentification et chiffrement de ce protocole.

Une URL (pour *Uniform Resource Locator*, signifiant littéralement « localisateur de ressource uniforme ») est une chaîne de caractères décrivant l'emplacement d'une ressource. Elle contient généralement dans l'ordre<sup>7</sup> : le nom du protocole, un deux-points (:), deux barres obliques (//), un nom d'hôte, une barre oblique (/), et un chemin composé de texte séparés par des barres obliques. Exemple : `http://example.com/un/chemin/page.html`.

Un hyperlien (ou dans le langage courant, un lien) est un élément dans une ressource qui est 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. Ainsi, il est possible d'établir un hyperlien sans la moindre intervention, ni coopération, de la ressource cible de l'hyperlien. Ce modèle unilatéral adopté par le Web ne se retrouve pas dans tous les systèmes hypertextes. Il a l'avantage prépondérant, dans un réseau public mondial, de permettre la création d'innombrables hyperliens entre partenaires décentralisés et non coordonnés.

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.

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 seul hôte peut contenir les deux.

## **Serveur HTTP**

---

Un serveur HTTP (ou service web) est un logiciel utilisé pour servir des ressources à travers le protocole HTTP. Un client HTTP est un logiciel utilisé pour manipuler ces ressources.

Un serveur web est un hôte sur lequel fonctionne un serveur HTTP (ou service web). Un serveur web peut héberger les ressources qu'il dessert, peut les récupérer sur des hôtes distants, délivrer ces ressources telles quels, on parlera alors de ressources statiques), ou bien les modifier en fonction de différents critères, on parlera alors de ressources dynamiques. Il est courant sur les services de grosse taille que les ressources soient réparties sur différents serveurs web et

éventuellement différents types de serveurs HTTP, certains ayant de meilleures performances pour les fichiers statiques, d'autre plus de souplesse pour délivrer des fichiers dynamiques. Un serveur web peut être un hôte spécialisé, une partie d'une infrastructure spécialisée ou bien encore un hôte plus généraliste. On appelle site web, l'ensemble des ressources HTTP associées à un nom de domaine et un ou plusieurs noms d'hôtes. 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.

## **Client HTTP**

---

Un client HTTP est un outil permettant d'utiliser les ressources fournies par un serveur HTTP. Il est chargé en relation avec différents services, de faire traduire le nom d'hôte en adresse IP par un DNS, de rentrer en contact avec le serveur HTTP, et de lui fournir différentes informations telles que le nom du site, et la ressource désirée, ainsi que ses capacités (gestion de différents types de ressource, capacité d'affichage, possibilité de compression de données, etc.) ou sa volonté (choix des langues, protection de la vie privée) dans l'utilisation des données fournies.

Un navigateur web est un type de logiciel client HTTP à interface homme-machine et 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 ressources et d'autres protocoles de communication dépend du type de navigateur.

Un robot d'indexation est, dans le domaine du web, un type de logiciel client HTTP, utilisé pour indexer les différentes ressources de sites web.

Un agrégateur est, dans le domaine du web, un type de logiciel client HTTP, permettant de regrouper les fils de syndication de différents sites web.

Un aspirateur de site web est un client HTTP permettant de récupérer l'intégralité d'un site web, pour le consulter ensuite hors-ligne ou en conserver une archive.

## **Documents**

---

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 ou XHTML) et d'éventuelles ressources liées, automatiquement accessibles (par exemple, des images, des feuilles de style, des scripts javascript ou des fontes de caractères). Il n'est pas exceptionnel qu'un seul document HTML fasse appel à entre quelques dizaines à une centaine de ressources annexes.

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.

Un CMS (Content Management System) est un moteur de site, généralement spécialisé pour un site, permettant d'éditer facilement tout ou une partie du contenu du site.

## **Consultation et mesure**

---

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 (visiteur ou internaute).

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.

On appelle audience, la fréquentation d'un site.

On appelle fil d'Ariane, le parcours effectué par un visiteur sur un site.

On appelle journal d'accès, la liste des accès effectués par les différents utilisateurs sur un serveur HTTP.

On appelle référent, la ressource ayant conduit un visiteur à une ressource servie. Elles sont données par défaut par la majorité des clients HTTP, mais peuvent être bloquées. Le référent est généralement également conservé dans le journal du serveur HTTP en association avec la ressource consultée.

On appelle User-Agent, la référence donnée par le client HTTP au serveur HTTP pour indiquer sa nature. Parmi les informations données, il peut y avoir le nom du client HTTP et sa version, le nom du moteur de rendu HTML utilisé, les langues préférées par l'utilisateur ou l'outil client, ainsi que différents autres critères comme les modules d'extension installés sur le client. Ces critères permettent au serveur HTTP d'évaluer quelles ressources seront plus adaptées à celui-ci. Différentes technologies dynamiques comme javascript ou les CSS, permettent également aux pages de supporter des clients hétérogènes et de s'adapter dynamiquement en utilisant ses ressources plutôt que celles du serveur.

La mesure de l'audience est l'étude des consultations effectuées sur un site, elle vise à comprendre les désirs des utilisateurs, en fonction des ressources les plus utilisées et du fil d'Ariane de ceux-ci.

La mesure d'audience peut se faire de différentes manières :

- Analyse des journaux du serveur. Les calculs de mesure de l'audience peuvent alors être effectués en temps réel au moment de la visite, ou bien en différé.
- Utilisation de scripts sur le serveur web lui-même.
- Utilisation d'images invisibles déportées sur un serveur web tiers (mesure d'audience simple).
- Utilisation de script en langage javascript et envoi des informations au serveur web ou a un serveur tiers (mesure d'audience plus complète).

## **Métiers du web**

---

On appelle hébergeur web une personne physique ou morale hébergeant (rendant accessible en ligne) sur son ou ses serveurs, les ressources constituant les sites web de ses clients. Cela peut aller du simple auto-hébergement d'un particulier à celui d'une association, d'une entreprise ou d'un industriel. L'hébergement peut alors être fait à domicile sur un modem, assurant un minimum de fiabilité, dans pièce climatisée dédiés aux serveurs dans un local généraliste au centre de données spécialisé, assurant la redondance énergétique et réseau nécessaire au fonctionnement sans interruptions 24/7.

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

## **Types de services web**

---

- Un service web est une technologie client-serveur fondée sur les protocoles du web.
- Un annuaire web est un site web répertoriant des sites web.
- Un moteur de recherche est un site permettant d'effectuer des recherches sur l'ensemble des sites web selon différents critères.
- 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 les met en forme de façon ergonomique et/ou attractive.
- Un blog est un journal web, fournissant des informations dépendantes du temps. Un blog BD est un type de blog utilisant le medium de la bande-dessinée comme support de communication.
- Un forum est un site permettant de faire des échanges entre différents utilisateurs, utilisé notamment dans le support ou la communication.
- Un site d'e-commerce est une boutique en ligne permettant de commander des produits ou services à distance.
- Un comparateur est un site chargé de comparer différents produits, services ou prix.

- Un gestionnaire de projet, est un outil de projet et de support pour différents types d'activité.
- Un Logiciel de suivi de problèmes est un site spécialisé dans le support, il permet de trier les problèmes de retrouver des traces de problèmes et les solutions apportées et d'échanger des données permettant à la résolution.
- Un webmail est un lecteur de courrier électronique utilisant une interface web.
- Un wiki est un site web éditable par les utilisateurs.

## Histoire



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

Tim Berners-Lee (TBL) travaille comme informaticien dans le bâtiment 31 (🌐46° 13' 57" N 6° 02' 42" E<sup>8</sup>) 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 CERN<sup>9</sup>. 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 Internet<sup>10</sup>. 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 promotion<sup>11</sup>. 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, créeront 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 30 avril 1993, le CERN verse dans le domaine public toutes les technologies développées autour du www<sup>12</sup>. Pour l'occasion, l'organisation place en ligne ce qui peut être considéré comme étant le premier site en "www"<sup>13</sup>, hébergé sur la station NeXT de TBL<sup>12</sup>.

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. 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.

## Chronologie

---

Les premières années de cet historique sont largement issues de *A Little History of the World Wide Web (Une petite histoire du World Wide Web)*.

- 1989
  - Le 13 mars, Tim Berners-Lee, engagé au CERN à Genève en 1984 pour travailler sur l'acquisition et le traitement des données<sup>14</sup>, propose de développer un système hypertexte organisé en *Web*, afin d'améliorer la diffusion des informations internes : *Information Management: A Proposal*<sup>9</sup>.
- 1990



Le premier serveur web, un NeXT Cube

- Robert Cailliau rejoint le projet et collabore à la révision de la proposition : *WorldWideWeb: Proposal for a HyperText Project*<sup>3</sup>.
- **Étendue** : Le premier serveur web est `nxoc01.cern.ch` ; la première page web est `http://nxoc01.cern.ch/hypertext/WWW/TheProject.html` ; la plus ancienne page conservée date du 13 novembre.
- **Logiciels** : Le premier navigateur web, appelé WorldWideWeb (plus tard rebaptisé Nexus) est développé en Objective C sur NeXT[2]. En plus d'être un navigateur, *WorldWideWeb* est un éditeur web. Le navigateur mode

texte line-mode est développé en langage C pour être portable sur les nombreux modèles d'ordinateurs et simples terminaux de l'époque.

- **Technologies** : Les trois technologies à la base du web, URL, HTML et HTTP, sont à l'œuvre. Sur NeXT, des feuilles de styles simples sont également utilisées, ce qui ne sera plus le cas jusqu'à l'apparition des Cascading Style Sheets.
- 1991
  - Le 6 août, Tim Berners-Lee rend le projet *WorldWideWeb* public dans un message sur Usenet [3].
  - **Étendue** : premier serveur web hors d'Europe au SLAC ; passerelle avec WAIS [4].
  - **Logiciels** : fichiers développés au CERN disponibles par FTP.
- 1992
  - Le *World Wide Web* fait l'objet d'une présentation d'ensemble datée du 3 novembre 1992<sup>15</sup>.
  - **Étendue** : 26 sites web « raisonnablement fiables » [5].
  - **Logiciels** : navigateurs Erwise [6], ViolaWWW [7] ; serveur NCSA HTTPd.
- 1993
  - Le 30 avril, le CERN renonce aux droits d'auteur sur les logiciels du *World Wide Web* et permet leur mise dans le domaine public [8]. À la fin de l'année, les médias grand public remarquent Internet et le WWW.
  - **Étendue** : 130 sites web en juin, 623 en décembre [9] ; l'usage croît d'un rythme annuel de 341 634 %.
  - **Logiciels** : Apparitions des navigateurs NCSA Mosaic et Lynx. Disponible d'abord sur X Window, puis sur Windows et MacOS, Mosaic cause un phénoménal accroissement de la popularité du web.
  - **Technologies** : images dans les pages web (Mosaic 0.10) ; formulaires interactifs (Mosaic 2.0pre5).
- 1994
  - **Étendue** : 2 738 sites en juin, 10 022 en décembre.
  - **Sites** : Yahoo! créé par deux étudiants ; apparition de la publicité sur HotWired.
  - **Logiciels** : Netscape Navigator 1.0.
  - **Standards** : fondation du World Wide Web Consortium ; RFC 1738 (*Uniform Resource Locators*).
- 1995
  - Microsoft crée MSN pour concurrencer Internet et le web, puis change d'avis et lance la guerre des navigateurs.
  - **Étendue** : 23 500 sites en juin (18 957 en août selon la première mesure de Netcraft [10]).
  - **Logiciels** : serveur HTTP Apache ; Microsoft Internet Explorer 1.0 et 2.0.
  - **Sites** : moteur de recherche AltaVista.

- **Technologies** : formatage tabulaire (Netscape Navigator 1.1b1), documents multi-cadres (Netscape Navigator 2.0b1), Java, JavaScript (Netscape Navigator 2.0b3), PHP.
- **Standards** : RFC 1866 (HTML 2.0).
- 1996
  - **Étendue** : 100 000 sites en janvier, environ 230 000 en juin.
  - **Logiciels** : Netscape Navigator 2.0 et 3.0 ; Internet Explorer 3.0 ; Opera 2.1.
  - **Standards** : RFC 1945 (HTTP/1.0) ; CSS level 1.
  - **Sites** : Internet Archive commence à archiver le web.
- 1997
  - **Étendue** : plus de 1 000 000 sites en avril selon Netcraft.
  - **Logiciels** : Netscape Navigator 4.0 ; Internet Explorer 4.0.
  - **Standards** : HTML 3.2 ; HTML 4.0.
- 1998
  - America Online rachète Netscape qui a perdu la guerre des navigateurs pour 4 milliards US\$.
  - **Étendue** : plus de 2 000 000 sites en mars.
  - **Logiciels** : Netscape Navigator 4.5 ; projet Mozilla.
  - **Sites** : Google.
  - **Standards** : XML 1.0 ; CSS level 2 ; DOM level 1, WAP level 1.
- 1999
  - **Étendue** : plus de 4 000 000 de sites en janvier, plus de 7 400 000 en août.
  - **Logiciels** : Internet Explorer 5.0.
  - **Standards** : HTML 4.01 ; RFC 2616 (HTTP/1.1).
- 2000
  - **Étendue** : 11 161 854 sites en février, 19 823 296 en août selon Netcraft [11].
  - **Standards** : XHTML 1.0.
- 2001
  - **Étendue** : 27 585 719 sites en janvier, 30 775 624 en août.
  - **Logiciels** : Internet Explorer 6.
  - **Sites** : Wikipédia.
- 2002
  - **Étendue** : 36 689 008 sites en janvier, 35 991 815 en août.
  - **Logiciels** : Mozilla 1.0.
- 2003
  - **Étendue** : 35 863 952 sites en février, 42 807 275 en août.
  - **Logiciels** : Safari.
- 2004
  - Le concept de Web 2.0 (désignant un phénomène pré-existant) apparaît.
  - **Étendue** : 46 067 743 sites en janvier, 53 341 867 en août.

- **Standards** : création du Web Hypertext Application Technology Working Group (WHATWG).
  - **Sites** : Facebook
  - **Logiciels** : Mozilla Firefox 1.0.
  - 2005
    - **Étendue** : 59 100 880 sites en février, 70 392 567 en août
    - **Logiciels** : Mozilla Firefox 1.5.
  - 2006
    - **Étendue** : 76 184 000 sites en février, 92 615 362 en août<sup>[réf. nécessaire]</sup>.
    - **Logiciels** : Internet Explorer 7, Mozilla Firefox 2.0.
  - 2007
    - **Étendue** : 108 810 358 sites en février<sup>16</sup>, 127 961 479 en août<sup>17</sup>.
    - **Logiciels** : Safari sur Windows.
    - **Standards** : Le W3C décide de poursuivre le travail sur HTML5 commencé au WHATWG.
  - 2008
    - **Étendue** : 158 209 426 sites en février<sup>18</sup>, 176 748 506 en août<sup>19</sup>.
    - **Logiciels** : Mozilla Firefox 3.0 ; Google Chrome 0.2.
  - 2009
    - **Étendue** : 216 000 000 sites en février<sup>20</sup>
    - **Logiciels** : Internet Explorer 8 ; Mozilla Firefox 3.5, qui inclut TraceMonkey, premier moteur d'exécution JavaScript avec compilation à la volée ; Safari 4 ; Opera 10
  - 2010
    - **Logiciels** : Mozilla Firefox 3.6 ; Google Chrome 4 ; Safari 5
  - 2011
    - **Étendue** : 284 842 077 sites en février<sup>21</sup> 312 693 296 sites en avril<sup>22</sup>
    - **Logiciels** : Google Chrome 9 (février), 10 (mars), 11 (avril), 12 (juin), 13 (août), 14 (septembre), 15 (octobre), 16 (décembre) ; Mozilla Firefox 4 (mars), 5 (fin juin), 6 (fin août) et 7 (fin septembre) ; Internet Explorer 9 ; Opera 11
  - 2012
    - **Logiciels** : Mozilla Firefox 16 (octobre) ; Google Chrome 19 (mai) ; Safari 6.0 (juillet)
  - 2013
    - Dans un projet de restauration historique<sup>23</sup>, le CERN remet en ligne <http://info.cern.ch/hypertext/WWW/TheProject.html>.
    - **Logiciels** : Internet Explorer 11
  - 2014
    - **Étendue** : 1 000 000 000 sites en septembre<sup>24</sup>
    - **Logiciels** : Mozilla Firefox 29 ;
-

## 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 collaboration<sup>25</sup>.

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.

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'Altavista<sup>note 2</sup> 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.

## **HTML5**

---

En 2007 le W3C reprend la technologie HTML5 et projette de le normaliser à l'horizon 2014. Dès l'année 2012, la majorité des moteurs de rendu HTML permettent d'exploiter ce standard et en 2013, la majorité des Systèmes de gestion de contenu utilisent une base HTML5. HTML5 apporte la possibilité d'utiliser des vidéos et de l'audio dans les pages sans dépendre d'un format propriétaire, avec les normes MPEG-4, WebM et Ogg Vorbis, d'utiliser davantage de graphiques vectoriels avec l'utilisation de la norme SVG, de faire de la visioconférence avec webRTC<sup>26</sup>. Le développement de WebGL permet désormais d'utiliser des objets en trois dimensions dans les applications web avec une grande interopérabilité. Le format JSON permet d'échanger des données, en relations avec des bases ou des systèmes de fichiers, augmentant au passage les possibilités de référencement, et d'automatisation des tâches entre les applications d'Internet.

## **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.

## **Répartition du trafic web**

---

Selon une étude de la société Incapsula réalisée en mars 2013, 50% du trafic web était généré par des bots, c'est-à-dire des machines automatisées. Une mise à jour récente de cette étude (décembre 2013)<sup>27</sup> révèle que ces mêmes machines représentent aujourd'hui près de 61% du trafic avec notamment une hausse du trafic des moteurs de recherche en hausse de 55%. Les opérations de hacking représentent environ 4.5% du trafic alors que les humains prennent environ 38.5%. Autrement dit à peine plus du tiers du trafic est généré par les activités humaines. La cause réside probablement dans l'explosion du marketing qui représente à lui seul la moitié du trafic (bots et humains confondus).

---

# 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-sac<sup>28</sup> : 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 invariants d'échelle 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 web<sup>29</sup>. 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

---



### 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 décompression 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 a 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 standardisées 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 fichiers du serveur ou un système de fichiers 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. Le premier serveur HTTP fut CERN httpd et est rapidement tombé en désuétude. Aujourd'hui<sup>[Quand ?]</sup>, les principaux serveurs Web utilisés sont Apache HTTP Server, serveur dominant du marché depuis une dizaine d'années, Nginx utilisé sur les sites à plus forte audience, mais aussi Microsoft IIS, Google Web Server ou encore Lighttpd.

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 contrôler le nombre d'instances des applications dynamiques du côté serveur.

Certains langages couramment utilisés avec les services 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, il y a quelques années, via mod fastcgi et aujourd'hui davantage via mod fcgid, soit via php-fpm.

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 2010<sup>30</sup>. 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 2010<sup>30</sup>.

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 2010<sup>30</sup>. Le langage Ruby, avec principalement le serveur d'application Ruby on Rails avec 0,5 % en 2010<sup>30</sup>. Devant la popularité grandissante de JavaScript côté 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.

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.