

Séance 6

Documents pour le web

Histoire et présentation du web

Internet, ou internet, cela n'a pas grande importance, est un réseau de réseaux, constitué par des ordinateurs connectés les uns aux autres. Sa structure est maillée, non pyramidale. Il y a bien des échelons, mais les échelons les plus élevés sont, en fait, assez proches de la base, et surtout, en très grand nombre. Ces échelons sont interchangeable. Quand votre échelon supérieur sera absent, en panne ou disparu, vous pourrez utiliser un autre échelon de même type. Il est ainsi très difficile de « diriger » Internet. Comme tout bon processus de communication, Internet dispose d'un langage. Ce langage est commun à toutes les machines informatiques, et indifférents aux plates-formes (à des degrés divers). Les données provenant d'Internet sont donc, moyennant parfois quelques retraitements, réutilisables par tout ordinateur.

Ainsi, un site Web est construit indépendamment de la plate-forme sur laquelle on le conçoit. Ce sont les logiciels qui s'adaptent aux machines. Il existe un Firefox, un navigateur web, pour PC et une autre version pour Mac ou pour Linux. Les données qu'ils exploitent, par contre, sont indépendantes des systèmes d'exploitation (Windows, Mac OS, Linux, etc.).

L'une des composantes les plus connues de l'Internet est le Web (World Wide Web). L'Internet comporte bien d'autres services : le courrier électronique, le transfert de fichiers, etc. Une confusion vient du fait que parfois, les logiciels utilisés pour le Web, incorporent également des fonctionnalités de courrier électronique (Mozilla Mail par exemple), on a alors tendance à croire que les deux fonctions se regroupent. Ce qui est faux. Pour envoyer un e-mail, vous pouvez utiliser Eudora, Outlook ou Thunderbird, qui est uniquement un logiciel de courrier.

Les fichiers utilisés pour « construire » le Web, pour le faire vivre, sont des fichiers HTML, utilisés par des logiciels particuliers : les navigateurs Internet Explorer, Firefox, etc.). Tout comme Word, logiciel de traitement de texte, lit les fichiers au format « .doc », les navigateurs Internet lisent, principalement, des fichiers au format « html ». Ces fichiers, stockés un peu partout dans le monde sur des ordinateurs interconnectés, par les liens qui les unissent, par le système d'adressage universel et par leur langage commun, sont le Web. On utilise le terme de Web « toile » pour représenter le système d'interconnexion complexe qui unit tous les sites les uns aux autres (système reposant sur l'hypertexte).

Notez, pour compliquer les choses, qu'on trouve des sites *Web* qui permettent l'envoi d'e-mail (Hotmail, Caramail, etc.). Si on utilise le Web pour accéder à ces sites, il n'en reste pas moins que pour la gestion des mails, les sites utilisent de leur côté les protocoles habituels pour le courrier.

Le principe du HTML

Définition

Le HTML (« HyperText Mark-Up Language ») est un langage dit de « marquage » (de « structuration » ou de « balisage ») dont le rôle est de formaliser l'écriture d'un document avec des balises de formatage. Les balises permettent d'indiquer la façon dont doit être présenté le document et les liens qu'il établit avec d'autres documents.

Le langage HTML permet notamment la lecture de documents sur Internet à partir de machines différentes, grâce au protocole HTTP, permettant d'accéder via le réseau à des documents repérés par une adresse unique, appelée URL.

Les pages web sont généralement organisées autour d'une page d'accueil, jouant un point central dans la navigation à l'aide des liens hypertextes. Cet ensemble cohérent de pages web liées par des liens hypertextes et articulées autour d'une page d'accueil commune est appelée site web.

Le Web est ainsi une énorme archive vivante composée d'une myriade de sites web proposant des pages web pouvant contenir du texte mis en forme, des images, des sons, des vidéo, etc.

En plus du texte adressé à votre lecteur, il vous faudra inclure des instructions pour le browser de celui-ci. Ces instructions seront différenciées de votre texte par les signes < et > par exemple <html>. Ces "instructions" s'appellent des tags ou des balises. Quand vous écrirez les balises de votre page HTML, Il faudra toujours garder à l'esprit - qu'une balise marque une action pour le browser.

ce qu'il doit faire

- que les attributs précisent les modalités de cette action.

comment il doit le faire

Schéma de fonctionnement :

<http://ndlr.info/leweb.ppt>

Le HTML est une norme

Il est important de comprendre que le langage HTML est un standard, c'est-à-dire qu'il s'agit de recommandations publiées par un consortium international : le World Wide Web Consortium (W3C).

Les spécifications officielles du HTML décrivent donc les "instructions" HTML mais en aucun cas leur implémentation, c'est-à-dire leur traduction en programmes d'ordinateur, afin de permettre la consultation de pages web indépendamment du système d'exploitation ou de l'architecture de l'ordinateur.

Toutefois, aussi étoffées les spécifications soient-elles, il existe toujours une marge d'interprétation de la part des navigateurs, ce qui explique qu'une même page web puisse s'afficher différemment d'un navigateur Internet à l'autre.

De plus, il arrive parfois que certains éditeurs de logiciels ajoutent des instructions HTML propriétaires, c'est-à-dire ne faisant pas partie des spécifications du W3C.

Ainsi des pages web contenant ce type d'instruction pourront être parfaitement affichées sur un navigateur et seront totalement ou en partie illisibles sur les autres, d'où la nécessité de créer des pages web respectant les recommandations du W3C afin de permettre leur consultation par le plus grand nombre.

PHP, ASP, JSP, etc.

Le format HTML n'est pas le seul reconnu par un serveur web. D'autres langages peuvent être utilisés : le PHP, l'ASP, le JSP, etc.

Le PHP est un langage de script qui permet d'inclure des petits programmes au sein d'une page web. Bien souvent on utilise le PHP pour interroger des bases de données. Lors de l'interprétation d'une page le code PHP est interprété par le serveur et la requête est envoyée à la base de donnée. C'est son résultat qui est présenté dans la page.

Un serveur web peut "présenter" tout type de fichier. Seuls certains peuvent être compris et interprétés par un serveur web.

Le cas du XML

XML (entendez eXtensible Markup Language et traduisez Langage à balises étendu, ou Langage à balises extensible) est en quelque sorte un langage HTML amélioré permettant de définir de nouvelles balises. Il s'agit effectivement d'un langage permettant de mettre en forme des documents grâce à des balises (markup). Contrairement à HTML, qui est à considérer comme un langage défini et figé (avec un nombre de balises limité), XML peut être considéré comme un métalangage permettant de définir d'autres langages, c'est-à-dire définir de nouvelles balises permettant de décrire la présentation d'un texte (Qui n'a jamais désiré une balise qui n'existait pas ?). La force de XML réside dans sa capacité à pouvoir décrire n'importe quel domaine de données grâce à son extensibilité. Il va permettre de structurer, poser le vocabulaire et la syntaxe des données qu'il va contenir.

En réalité les balises XML décrivent le contenu plutôt que la présentation (contrairement à HTML). Ainsi, XML permet de séparer le contenu de la présentation .. ce qui permet par exemple d'afficher un même document sur des applications ou des périphériques différents sans pour autant nécessiter de créer autant de versions du document que l'on nécessite de représentations !

XML a été mis au point par le XML Working Group sous l'égide du World Wide Web Consortium (W3C) dès 1996. Depuis le 10 février 1998, les spécifications XML 1.0 ont été reconnues comme recommandations par le W3C, ce qui en fait un langage reconnu. (Tous les documents liés à la norme XML sont consultables et téléchargeables sur le site web du W3C, <http://www.w3.org/XML/>)

XML est un sous ensemble de SGML (Standard Generalized Markup Language), défini par le standard ISO8879 en 1986, utilisé dans le milieu de la Gestion Electronique Documentaire (GED). XML reprend la majeure partie des fonctionnalités de SGML, il s'agit donc d'une simplification de SGML afin de le rendre utilisable sur le web.

Mise en page de XML

XML est un format de description des données et non de leur représentation, comme c'est le cas avec HTML. La mise en page des données est assurée par un langage de mise en page tiers. A l'heure actuelle (fin de l'année 2000) il existe trois solutions pour mettre en forme un document XML :

CSS (Cascading StyleSheet), la solution la plus utilisée actuellement, étant donné qu'il s'agit d'un standard qui a déjà fait ses preuves avec HTML

XSL (eXtensible StyleSheet Language), un langage de feuilles de style extensible développé spécialement pour XML. Toutefois, ce nouveau langage n'est pas reconnu pour l'instant comme un standard officiel

XSLT (eXtensible StyleSheet Language Transformation). Il s'agit d'une recommandation W3C du 16 novembre 1999, permettant de transformer un document XML en document HTML accompagné de feuilles de style

Comment produire du HTML

Voici vos premières balises ou tags :

<HTML> Ceci est le début d'un document de type HTML.

</HTML> Ceci est la fin d'un document de type HTML.

<HEAD> Ceci est le début de la zone d'en-tête.(Prologue au document proprement dit contenant des informations destinées au browser)

</HEAD> Ceci est la fin de la zone d'en-tête.

<TITLE> Ceci est le début du titre de la page.

`</TITLE>` Ceci est la fin du titre de la page.
`<BODY>` Ceci est le début du document proprement dit.
`</BODY>` Ceci est la fin du document proprement dit.
Vous aurez remarqué qu'à chaque balise de début d'une action, soit `<...>`, correspond (en toute logique) une balise de fin d'une action `</...>`.

Vous noterez aussi que les balises ne sont pas "case sensitive". Il est donc équivalent d'écrire `<HTML>`, `<html>`, `<Html>`, etc.

Voici comment créer un premier document HTML

Ouvrez l'éditeur de texte.
Ecrivez les codes Html suivants:

```
<HTML>
<HEAD>
<TITLE>Mon premier document web</TITLE>
</HEAD>
<BODY>
</BODY>
</HTML>
```

Enregistrer le document avec l'extension `.html` ou `.htm`. Donnez le nom que vous voulez au fichier.

Ouvrez le navigateur. Affichez le document via le menu Fichier / Ouvrir

Vous voyez le premier document Html.

Celui-ci est vide (et c'est normal) mais tout à fait opérationnel. Il faudra maintenant lui fournir votre information à l'intérieur des balises `<BODY></BODY>`. Remarquez que votre "TITLE" est présent dans la fenêtre de Netscape.

Pour vos éventuelles modifications, il n'est pas nécessaire de rouvrir à chaque fois le navigateur.

Retournez dans l'éditeur de texte (sans fermer le navigateur).

Modifiez les codes Html.
Enregistrez le fichier.

Utilisez la commande Reload du navigateur ou si celui-ci est paresseux cliquer dans la barre "Adresse" et pressez la touche "Entrée".

Utilisez le référentiel des différentes balises pour créer d'autres liens.

Mettre la nouvelle page en ligne en utilisant le FTP.

L'intérêt d'un format comme le PDF

Développé par Adobe, le format PDF (Portable Document Format) permet de créer des documents illustrés de plusieurs page pour un poids réduit. Ce format se révèle

particulièrement utile sur l'Internet. Outre la légèreté des documents qu'il produit, le PDF a aussi pour lui un atout sécurité : contrairement à des documents Microsoft Word qui peuvent véhiculer des virus macro, le PDF est un document "figé" qui se rapproche plus du fichier image que du document éditable. Adobe fournit toute une suite logicielle comprenant éditeurs et plugiciels pour créer et manipuler des PDF. Il existe également des solutions alternatives, gratuites, qui vous permettent de créer vos propres documents PDF. C'est le cas de PDFCreator.

PDFCreator n'est pas un éditeur de texte ou bien un convertisseur comme on peut en rencontrer dans le domaine de la musique (de WAV vers MP3, MP3 vers AAC, etc.). En fait, PDFCreator a ajouté une imprimante à votre ordinateur : une imprimante PostScript. C'est à partir de ce format de fichier que PDFCreator pourra produire un PDF à proprement parler. PostScript est un langage de description de page créé par Adobe en 1984 très utilisé par les imprimantes laser. Les caractères composant la page sont décrits par des courbes de Bézier. Il est ainsi possible d'enregistrer le résultat d'une impression PostScript dans un fichier qui aura l'extension *.ps*. C'est grâce à ce fichier que PDFCreator pourra produire un PDF.

Tous ces éléments montrent bien l'intérêt d'un système comme celui des blogs ou des sites dynamiques gérés avec PHP et base de données.

Les principales balises HTML

Mise en forme des caractères

<code>...</code>	Texte en gras
<code>...</code>	Texte en italique
<code>...</code>	Texte en couleur où XXXXXX est une valeur hexadécimale
<code>...</code>	Taille des caractères où X est une valeur de 1 à 7
<code><I>...</I></code>	Texte en italique
<code><SMALL>...</SMALL></code>	Réduction de la taille des caractères
<code>...</code>	Mise en gras du texte
<code><U>...</U></code>	Texte souligné

Mise en forme du texte

<code><!--...--></code>	Commentaire ignoré par le navigateur
<code>
</code>	A la ligne
<code><BLOCKQUOTE>...</BLOCKQUOTE></code>	Citation (introduit un retrait du texte)
<code><CENTER>...</CENTER></code>	Centre tout élément compris dans le tag
<code><DIV align=center> ...</DIV></code>	Centre l'élément encadré par le tag
<code><DIV align=left> ...</DIV></code>	Aligne l'élément à gauche
<code><DIV align=right> ...</DIV></code>	Aligne l'élément à droite
<code><P>...</P></code>	Nouveau paragraphe
<code><P align=center>...</P></code>	Paragraphe centré
<code><P align=left>...</P></code>	Paragraphe aligné à gauche
<code><P align=right>...</P></code>	Paragraphe aligné à droite

Listes

<code></code>	Liste non numérotée (dite à puces)
<code></code>	Élément de liste
<code></code>	
<code></code>	Liste numérotée
<code></code>	Élément de liste
<code></code>	
<code><DL></code>	Liste de glossaire
<code><DT>...</DT></code>	Terme de glossaire (sans retrait)
<code><DD>...</DD></code>	Explication du terme (avec retrait)
<code></DL></code>	

Ligne de séparation

<code><HR></code>	Trait horizontal (centré par défaut)
<code><HR width="x%"></code>	Largeur du trait en %
<code><HR width=x></code>	Largeur du trait en pixels
<code><HR size=x></code>	Hauteur du trait en pixels

<code><HR align=center></code>	Trait centré (défaut)
<code><HR align=left></code>	Trait aligné à gauche
<code><HR align=right></code>	Trait aligné à droite
<code><HR noshade></code>	Trait sans effet d'ombrage

Hyperliens

<code>...</code>	Lien vers une page Web
<code>...</code>	Lien vers une adresse eMail
<code>...</code>	Lien vers la page locale fichier.htm située dans le même répertoire
<code>...</code>	Définition d'une ancre
<code>...</code>	Lien vers une ancre
<code>...</code>	

Images

<code></code>	Insertion d'une image au format Gif ou Jpg (voir liens pour l'adressage)
<code></code>	
<code></code>	Mise à l'échelle de l'image en pixels (a comme effet d'accélérer l'affichage de la page)
<code></code>	Définition de la bordure d'une image avec lien
<code></code>	Texte alternatif lorsque l'image n'est pas affichée
<code></code>	Aligne l'image en bas
<code></code>	Aligne l'image au milieu
<code></code>	Aligne l'image en haut
<code></code>	Aligne l'image à gauche
<code></code>	Aligne l'image à droite
<code></code>	Espacement horizontal entre l'image et le texte
<code></code>	Espacement vertical entre l'image et le texte

Tableau

<code><TABLE>...</TABLE></code>	Définition d'un tableau
<code><TABLE width="x%"></code>	Largeur du tableau en %
<code><TABLE width=x></code>	Largeur du tableau en pixels
<code><TABLE border=x></code>	Largeur de la bordure
<code><TABLE cellpadding=x></code>	Espace entre la bordure et le texte
<code><TABLE cellspacing=x></code>	Épaisseur du trait entre les cellules
<code><TR>...</TR></code>	Ligne du tableau
<code><TD>...</TD></code>	Cellule du tableau
<code><TD bgcolor="#XXXXXX"></code>	Couleur d'une cellule de tableau
<code><TD width="x%"></code>	Largeur de colonne en %
<code><TD width=x></code>	Largeur de colonne en pixels
<code><TD align=center></code>	Texte dans la cellule centré
<code><TD align=left></code>	Texte dans la cellule aligné à gauche

<TD align=right>	Texte dans la cellule aligné à droite
<TD valign=bottom>	Alignement vers le bas du contenu d'une cellule
<TD valign=middle>	Alignement vers le haut du contenu d'une cellule
<TD valign=top>	Centrage vertical du contenu d'une cellule
<TD colspan=x>	Nombre de cellules à fusionner horizontalement
<TD rowspan=x>	Nombre de cellules à fusionner verticalement

Fichier Html

<HTML>...</HTML>	Début et fin de fichier Html
<HEAD>...</HEAD>	Zone d'en-tête d'un fichier Html
<TITLE>...</TITLE>	Titre affiché par le browser (élément de HEAD)
<BODY>...</BODY>	Début et fin du corps du fichier Html
<BODY bgcolor="#XXXXXX">	Couleur d'arrière-plan (en hexadécimal)
<BODY background="xyz.gif">	Image d'arrière-plan

Le principe du site web :

- Les règles de base : donner son nom, son adresse, toujours donner la possibilité de revenir à la page d'accueil, etc.
- Qu'est-ce qu'on a le droit de publier ? De reproduire ? Peut-on faire un lien ? Comment choisir une licence pour son contenu web ?
- Comment faire connaître son site web ? Le principe et le fonctionnement des moteurs de recherche.
- Protéger l'accès à son site : retour au .htaccess

> Pratique

Suivre l'audience de son site, comprendre des statistiques

http://www.elec.ucl.ac.be/Stats/usage/usage_200510.html

Des statistiques permettent de connaître la fréquentation de son site, les pages d'entrée ou de sortie ou, même, quelles sont les pages qui ne sont jamais vues. Tout ce qu'il faut pour améliorer son site et connaître son audience.

Comprendre ses statistiques

Pour bien utiliser ses statistiques et en tirer pleinement parti, il faut comprendre les termes utilisés. Voici un bref glossaire.

Hits : représente le nombre total de requêtes faites au serveur (web) durant le laps de temps donné (mois, jour, heure, etc.).

Fichiers : représente le nombre total de hits (requêtes) qui engendreront un envoi en retour à l'utilisateur. Notez que tous les hits n'entraînent pas l'envoi de données (erreur 404 ou bien si la page demandée se trouve déjà dans le cache du navigateur).

Une grande différence entre hits et fichiers donne une indication du nombre de visiteurs qui reviennent sur votre site. De nombreux hits qui ne se transforment pas en fichiers transmis, cela peut vouloir dire que les pages sont déjà en cache, donc que les visiteurs sont déjà venus.

Sites : représente le nombre d'adresses IP ou de noms d'hôte uniques qui firent des requêtes sur votre serveur. Une adresse IP ou un nom d'hôte ne sont comptabilisés qu'une fois, même s'ils font plusieurs requêtes. Ce chiffre n'est pas le plus fiable pour ce qui est de comptabiliser vos visiteurs uniques. En effet, plusieurs visiteurs peuvent avoir la même adresse IP (connexion internet partagée ou proxy).

Visites : une visite représente en fait un ensemble. Elle commence lorsqu'une page est demandée sur votre serveur pour la première fois (pas dans l'absolu, dans le cadre de la visite, justement). Aussi longtemps que le même client (le même site) continue de faire des requêtes durant un certain laps de temps, il sera considéré comme une seule et unique visite. Par défaut, le temps de latence de la visite est de 30 minutes. Si deux requêtes sont espacées de plus de 30 minutes, la nouvelle requête commencera une nouvelle visite.

Seules les requêtes vers des pages seront prises en compte pour des visites. Les requêtes vers des images ou d'autres types de documents ne seront pas comptabilisées dans les visites (mais dans les hits et fichiers).

Page : une page consiste en un URL qui a une extension en .htm, .html ou .cgi, par exemple. Une page représente l'élément global appelé par une requête et non pas ses parties constituantes comme les images ou les fichiers audio.

Kbytes (KB) : représente 1024 bytes (ou 1 Kilobyte). Utilisé pour donner le volume de données qui ont été transférées entre le serveur et la machine cliente.

Site : une machine distante qui effectue une requête sur votre serveur. Pour définir un site, on se base sur son adresse IP et son nom d'hôte.

Référent (referrer) : un URL qui pointe vers votre site ou qui a amené un navigateur à demander un document sur votre serveur. La grande majorité de vos référents sera constituée de vos propres pages étant donné qu'elles pointent vers d'autres pages de votre site ou des documents présents sur votre site (des images, par exemple). Si l'une de vos pages contient des liens vers dix images présentes sur votre site, cela générera dix hits supplémentaires avec, comme référent, votre page web qui a appelé ces pages.

User Agents : User Agents est un autre nom pour browsers (navigateurs). Netscape, Opera, Konqueror, etc., sont tous des User Agents et chacun s'identifie d'une manière particulière auprès de votre serveur. Gardez toutefois à l'esprit que de nombreux navigateurs permettent à l'utilisateur de changer sa signature, sa méthode d'authentification. Vous risquez donc de trouver des noms fantaisistes dans vos statistiques. Pour des raisons de compatibilité, des navigateurs comme Konqueror ou Mozilla peuvent aussi signer comme Internet Explorer.

Entrée / Sortie : Correspondent respectivement à la première page appelée lors d'une visite (Entrée) et à la dernière page demandée (Sortie). Ces pages sont calculées en utilisant la logique de la visite. Lorsque d'une visite est démarrée, la page demandée est comptée comme une page d'entrée. Lorsque la visite se termine, le dernier URL demandé, quel qu'il soit, est enregistré comme page de Sortie.

Comment utiliser ses statistiques

En analysant vos statistiques vous serez en mesure de connaître la vie de votre site. Des différents chiffres et tableaux, quelques points sont particulièrement intéressants à analyser :

* les référents

Grâce à la liste des référents vous pouvez savoir d'où viennent vos visiteurs. Si vous avez investi plusieurs centaines d'euros dans des publicités et qu'aucun nouveau visiteur n'a une régie publicitaire ou un site sur lequel vous avez acheté de l'espace publicitaire en référent, vous pouvez en tirer deux hypothèses : soit vos bandeaux publicitaires ou vos annonces sont mauvais, soit les sites sur lesquels vous avez acheté de l'espace n'ont pas une audience suffisante ou leurs propres visiteurs ne sont pas concernés par votre activité.

C'est également grâce au référent que vous saurez si vos newsletters sont efficaces. Les webmails disposent d'un URL bien particulier.

* les pages d'entrée / sortie

Surveillez bien les pages d'entrée ainsi que les pages de sortie. Cela vous donnera une idée du parcours type d'un internaute sur votre site. Si la page de commande ou celle qui liste vos produits est la page de sortie la plus fréquente il vous faudra revoir la présentation de ces pages ou le déroulement du processus d'achat. Si les internautes sortent de votre site dès la page d'accueil c'est tout votre look qu'il faudra repenser (rendre la page plus claire, changer les accroches, etc.)

Les pages d'entrée peuvent vous aider à affiner votre politique commerciale. Si vous vendez des bateaux et des rames et que la majorité de vos visiteurs arrivent directement sur la page qui vend les rames, n'hésitez pas à mettre les rames plus en avant directement sur votre page d'accueil pour répondre plus rapidement aux attentes de vos visiteurs.

* visites

S'il est toujours agréable d'avoir un grand nombre de visiteurs cela ne suffit malheureusement pas à assurer le succès d'un site. Plus les visites sont longues, meilleur est le site. Cela veut dire que les internautes trouvent un contenu intéressant sur vos pages et qu'ils passent un certain temps à y surfer. Des visites trop courtes montrent que l'internaute se désintéresse vite des vos pages. Revoyez alors vos accroches ou la structure des pages pour les rendre plus claires.

Attention, des visites trop longues peuvent également être le signe que les internautes tournent en rond sur votre site sans réussir à y trouver l'information qu'ils recherchent. Surveillez alors les pages de sortie. Vous aurez peut-être à retravailler la page de sortie pour que son contenu y soit mieux présenté.