

Développement de Sites Web

Cours II :

Internet et HTML - une très brève introduction

Peter Stockinger

Séminaire de Maîtrise en Communication Interculturelle à l'Institut
National des Langues et Civilisations Orientales (INaLCO)

Paris, 2000 - 2002

Sommaire

Introduction.....	3
1) l'Internet et le Web.....	4
2) les balises html	8
3) l'architecture de base d'une page html	11
4) les commentaires.....	15
5) références.....	16

Introduction

Dans ce cours, seront introduits et présentés d'une manière très succincte les principales notions liées à l'Internet et au langage de balise HTML (Hypertext Markup Language) qui est indispensable pour l'édition et la "lecture" des pages sur le world wide web.

Il va de soi qu'il ne s'agit pas de produire un réel cours sur Internet - ni sur le HTML. Le seul objectif est de vous apporter quelques informations générales et tout à fait indispensables pour bien débiter ce stage de formation visant la production et la publication d'un site de recherche en communication interculturelle.

1) L'Internet et le Web

a) Internet

L'Internet est souvent appelé le "réseau des réseaux" (i.e. le réseau mondial qui intègre tous les autres réseaux - à condition qu'ils soient accessibles)

Un réseau est composé d'ordinateurs interconnectés, du matériel d'interconnexion et des technologies (logiciels) et standards rendant possible une inter-connexion.

On distingue entre différents types de réseaux :

- Réseau local,
- réseau physique,
- réseau "logique",
- etc.

L'Internet est composé de :

- **"backbones"** : lignes principales ("arêtes dorsales") reliant les ordinateurs des grands fournisseurs d'accès.
- Ces ordinateurs sont appelés **POP** (*Point of Presence* sur les lignes principales) :
- à ces POP sont reliés toute une diversité de petits **fournisseurs** qui, eux, vendent aux particuliers ou aux institutions publiques ou privées l'accès à l'Internet

b) Les protocoles :

Ce sont les "langages" de connexion, d'inter-connexion, i.e. rendant possible la connexion entre ordinateurs, l'envoi et la réception d'informations, etc.

TCP/IP (*Transmission Control Protocol/Internet Protocol*) : ils définissent les "rails", les "chemins" de transmission et assurent donc la transmission des informations d'un ordinateur à un ordinateur, voire entre différents ordinateurs.

SLIP (*Serial Line Internet Protocol*) : protocol permettant à un ordinateur individuel de se connecter par modem ou par RNIS à un serveur de son fournisseur ou à un POP

PPP (Point to Point Protocol) : amélioration du SLIP

c) des protocoles aux services:

Les protocoles sont à la base d'un ensemble de services Internet - services tels que :

Telnet : accès à distance aux ordinateurs (peu utilisé aujourd'hui)

FTP (File Transfert Protocol) : destiné à l'envoi et à la récupération de fichiers sur un ordinateur distant (très important pour le travail à distance)

NNTP (News Network Transfer Protocol) : accès aux newsgroups

SMTP (*Simple Mail Transport Protocol*), **POP2** (*Post Office Protocol*), **POP 3** : pour le courrier électronique

HTTP: (*Hypertext Transfert Protocol*) - protocole du world wide web : transfert de texte, d'images, de graphiques, ...

D'autres protocoles, plus spécialisés et aussi plus récents concernent le transfert sécurisé d'informations, la diffusion des contenus vidéo, etc.

d) L'adressage

Il est crucial que chaque ordinateur "participant" aux échanges sur Internet, possède - d'une manière analogue aux coordonnées d'une personne - sa propre (et unique) **adresse**. C'est le rôle du numéro IP :

Le numéro IP : chaque ordinateur possède son numéro IP

Numéro IP - 4 octets

Exemple : 193. 222. 566. 112 (numéro fictif)

Selon l'exploitation des numéros IP (composés de quatre octets), on reconnaît trois types de réseaux :

- **Réseaux de classe C** : on utilise uniquement le dernier octet pour l'adressage des ordinateurs du réseau (réseau est limité à 256 = 2 puissances 8 ordinateurs) : "petit réseau"
- **Réseaux de classe B** : on utilise les deux derniers octets pour l'adressage des postes - les deux premiers étant réservés à l'adressage du réseau lui-même (taille maximale de ce réseau : 65536 ordinateurs); il s'agit du réseau le plus répandu. Les réseaux de la MSH et de l'INALCO en font partie
- **Réseaux de classe A** : on utilise les trois derniers octets pour l'adressage des postes du réseau. Ce sont des réseaux assez gigantesques (Taille maximale : plus que 16 millions d'ordinateurs reliés) en service, par exemple, dans les grandes multinationales.

e) Le DNS (Domain Name System)

Le numéro IP désigne d'une manière univoque une machine. Mais on peut regrouper des IP sous un "**nom**", i.e. **nom de domaine**, qui désigne un ou un ensemble d'ordinateurs comme appartenant à un réseau (physique, logique, ...) donné.

Exemple :

Ordinateur (Serveur)	Domaine		Domaine de Haut Niveau
	"Local Level Domain"	Second Domain Level	
			com
www.		semionet.	
Semioweb.		msh-paris.	fr
semioweb	recherche.	msh-paris.	

Domaine de haut niveau :

"fr" : abréviations des noms de pays (.de, .at, .es, .it, ...)

"com" : entreprises, sociétés, ...

"edu" : organismes scolaires, universités, ...

"gov" : gouvernement, ...

"org" : organisations non-commerciales

etc.

f) URL (Uniform Resource Locator)

URL : toute les données sur Internet (documents, pages, programmes, ...) possèdent leur adresse précise sur Internet

Protocol	Hôte: port	Chemin	Fichier
http://	www.semionet.com/	Maitrise200 2/	projcoll2002.htm
http://	(fictif) 193.144.333.888 : 80/	Maitrise200 2/	projcoll2002.htm
ftp://	(fictif) 193.144.333.888 : 21/	Maitrise200 2/	coursII.pdf

g) Le Html

Le world wide web est constitué d'une gigantesque bibliothèque de ressources d'information - documents textuels, images, graphiques et dessins, vidéos, animations, programmes, données structurées, etc.

Une très grande partie de ces ressources d'information se présente, à l'heure actuelle, sous forme de "pages" - des "pages numériques" - qui sont codées dans le langage **HTML** (*Hypertext Markup Language*).

Le HTML est un langage à **balises**. Cela signifie qu'il existe un **corps de balises** (environ 100) et d'attributs (environ 150) qui sont plus ou moins généralement reconnues et respectées par les producteurs de logiciels de **production** et **d'exploration** (de "lecture") de pages web :

- Logiciels de production : **éditeurs HTML** tels que Dreamweaver de Macromedia, Frontpage de Microsoft, PageMill et GoLife d'Adobe, Namo, etc.;
- Logiciels de "lecture" : **"navigateurs"** tels que Internet Explorer de Microsoft, Netscape Communicator, Mozilla, Opera, ...)

Ainsi, (en principe..) peu importe le matériel (ordinateur, ..) et peu importe aussi le logiciel utilisés pour éditer ou lire une page web, n'importe qui se tient à ce langage et ses balises pourra communiquer, échanger des informations avec qui que ce soi à condition, de nouveau, que ce dernier accepte également les règles imposées par le HTML.

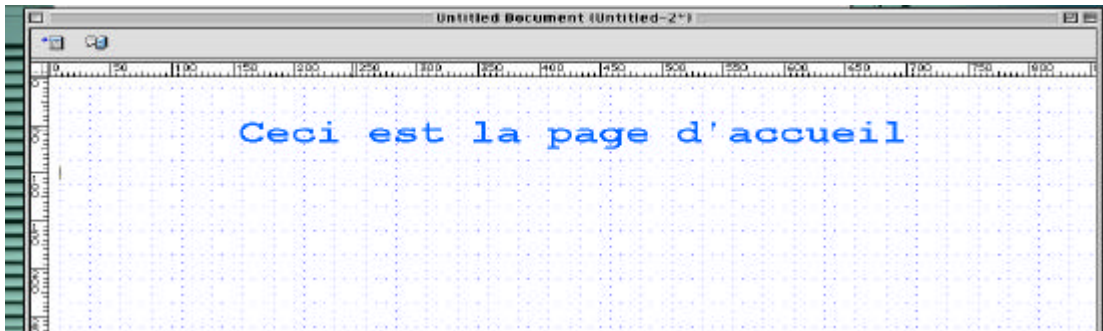
Ceci dit, chaque producteur (Netscape, Microsoft, ...) a développé, en dehors des balises définies comme "noyau" du langage HTML, ses propres balises ce qui rend parfois relativement délicate la question de l'interprétation correcte d'une page développée par tel ou tel éditeur dans un navigateur appartenant à un concurrent de ce dernier (c'est le cas dans l'histoire épique entre Microsoft et Netscape ...).

On connaît différentes versions du langage HTML. La version 4 est la plus récente et aussi la plus utilisée, à ce jour.

2) les balises html

Une balise (ou "tag") est une **instruction** (à un logiciel d'édition ou de lecture) de traiter une donnée qu'elle contient selon une façon définie.

Prenons l'exemple ci-dessous



Il s'agit d'un simple texte, qui occupe ici la fonction d'un *titre*. Le texte est rendu par la *police* Courier, en *couleur* bleue, etc. Il est *centré* au milieu de la page.

Ces informations (titre, police de caractère, couleur, alignement, ...) sont contenues dans les balises et les attributs des balises :



L'information "titre" est définie par la balise `<h1>` et l'information "police" est définie par la balise ``

Les deux balises sont spécifiées par des attributs (i.e. des propriétés qui leurs sont propres). Ainsi, la balise `<h1>` comporte l'attribut *align* et la balise `` contient les trois attributs *face*, *size* et *color*.

Un attribut, lui est composé par

- le **nom** de l'attribut (par exemple : "face" ou "size"),
- un **signe** d'égalité (" = ") et
- une **valeur** de l'attribut (par exemple "Courier", "+5" ou encore la formule bizarre #0066FF qui signifie une certaine teinte en bleu).

Autrement dit, l'expression :

`color = "#0066FF"`

signifie grosso modo que la couleur (de la police) est bleue. L'expression

`align = "center"`

signifie que le titre doit être centré (i.e. se trouver au milieu de la page).

Les deux balises identifiées ci-dessus avec leurs attributs (1 pour la balise titre; 3 pour la balise font) suffisent pour assurer :

- 1) une production (édition) correcte du texte "Ceci est la page d'accueil" sur le web à condition qu'on dispose d'un éditeur HTML (i.e. d'un éditeur tenant compte du langage HTML);
- 2) une lecture correcte du dit texte par n'importe quel utilisateur à condition qu'il possède un "navigateur" tenant compte du langage HTML

Il faut noter que toutes les balises (ou "tags") sont délimitées par des crochets (signe inférieur < et signe supérieur >). Cette structure est toujours la même. Cependant, chaque balise doit être ouverte et fermée (sauf rares exceptions). Voici donc l'écriture correcte d'une balise :

`<h1>..... </h1>`

Les majuscules et les minuscules dans les instructions n'ont pas importance, même s'il vaut mieux écrire toutes les balises en minuscules

L'attribut et sa valeur est inséré juste après l'ouverture et la déclaration de la balise à laquelle il appartient, cf:

`<h1 align="center">..... </h1>`

`..... `

Voici quelques familles principales de balises que l'on peut distinguer dans le langage html :

- Balises contenant des méta-informations (auteur, mots-clé, description, titre, ...)
- Balises de mise en forme du document,
- Balises pour la création de cadres,
- Balises de mise en forme du texte
- Balises de liens hypertextuels et hypermédias
- Balises pour formulaires
- Balises d'insertions des images, du son et de la vidéo
- Balises pour VRML et 3D
- Etc.

Pour avoir une idée plus précise sur le langage HTML (version 4) et les différentes balises + attributs, on peut consulter le petit livre de :

Equipe Sémiotique Cognitive et Nouveaux Médias (ESCoM)
Maison des Sciences de l'Homme (MSH)
54, Bd. Raspail - 75006 Paris - France

Ralph Steyer, *HTML 4 & HTML dynamique*. Paris, Micro Application 1998

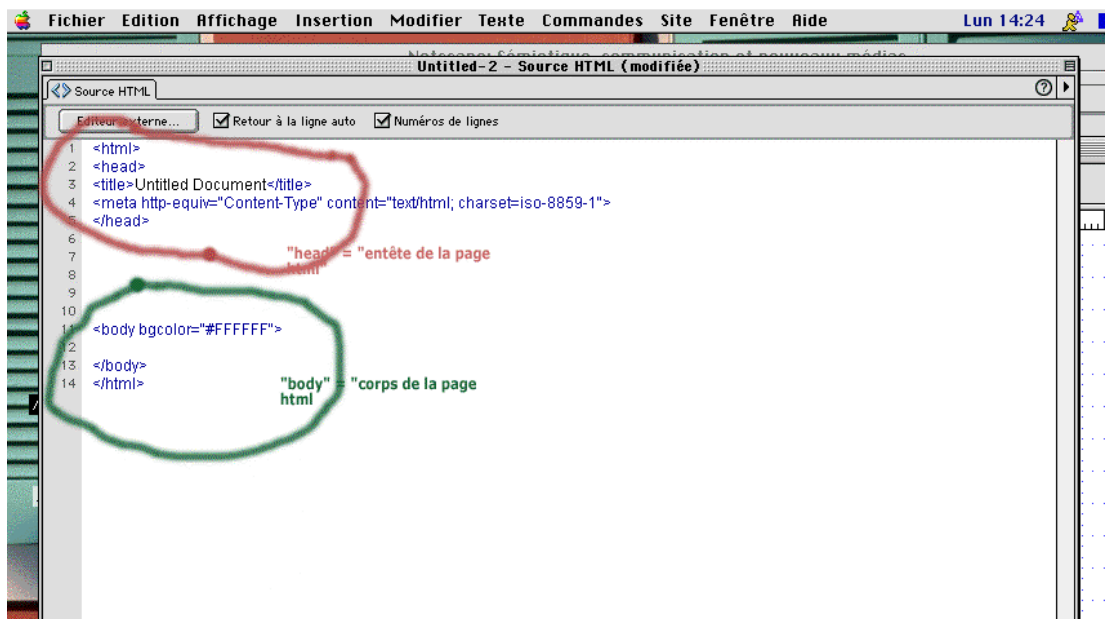
En effet, il n'est pas nécessaire de connaître toutes les balises. Il suffit de savoir comment le langage HTML fonctionne afin de pouvoir diagnostiquer d'éventuelles erreurs dans la production/affichage d'une page. Cela veut dire qu'il est très conseillé de consacrer quand même un petit moment à l'étude des balises html et de l'architecture d'une page web.

Il existe toute une gamme de logiciels sophistiqués d'édition de pages html qui permettent la création de pages web de très haut niveau avec un minimum de connaissances du langage html.

3) l'architecture de base d'une page html

D'une manière canonique, une page HTML est constituée de deux parties :

- Une partie appelée "tête" ("**head**", en anglais)
- Une partie appelée "corps" ("**body**", en anglais).



Dans la partie "tête"

Dans la partie "tête" d'une page html, on trouve les informations générales : titre, auteur, mots-clé, etc. cf. ci-après l'extrait des informations que l'on peut trouver dans la page d'accueil du site de l'ESCoM :

```
<html>
<head>
<meta http-equiv="Content-Type" content="text/html; charset=iso-8859-1">
<meta name="keywords" content="escom, sémiotique structurale, sémiotique textuelle, sémantique,
lexicologie, analyse du discours, genre textuel, communication, organisation sociale">
<meta name="description" content="Bienvenue sur le site de l'ESCoM - Equipe Sémiotique
Cognitive et Nouveaux Médias. ESCoM est un centre de recherche spécialisé dans la sémiotique
appliquée et les nouvelles technologies et situé à la Maison des Sciences de l'Homme à Paris, France).">
<title>Sémiotique, communication et nouveaux médias - ESCoM</title>
<meta name="auteur" content="This web site created by Elisabeth de Pablo,
is maintained by the researchers of ESCoM. For more information, please contact
contact@semionet.com">
<meta name="robots" content="index,follow">
<link href="http://semioweb.msh-paris.fr/escom" title="ESCoM">
</head>
```

Une information - très importante - est celle relative à l'usage des différents standards utilisés: par exemple le standard d'encodage des polices de caractères (important, par exemple, dans un contexte multilingue) - cf:

Equipe Sémiotique Cognitive et Nouveaux Médias (ESCoM)
Maison des Sciences de l'Homme (MSH)
54, Bd. Raspail - 75006 Paris - France

```
<head>
<meta http-equiv="Content-Type" content="text/html; charset=iso-8859-1">
</head>
```

L'entête peut être précédée par une déclaration de la version html utilisé pour l'édition de la page ...
cf:

```
<!DOCTYPE HTML PUBLIC "-//W3C//DTD HTML 4.0 Transitional//EN">
<!-- ***** -->
<!-- This Web site and Internet application is produced, -->
<!-- designed, programmed and maintained by -->
<!-- ESCOM -->
-->
<!-- (Equipe Semiotique Cognitive et Nouveaux Medias) -->
<!-- Tel: 00 33 1 49 54 21 83 / 22 30 / 22 54 -->
<!-- stock@msh-paris.fr -->
<!-- http://semioweb.msh-paris.fr/escom -->
<!-- ***** -->
<html>

<head>
<meta http-equiv=".....
.....
```

Enfin, dans l'entête on trouve également la déclaration de toute sorte de scripts (nécessaires pour la production de menus contextuels, de formulaires interactifs, de tests et jeux, etc.). Nous y reviendrons dans un cours ultérieur dédié aux aspects dits interactifs d'un site web.

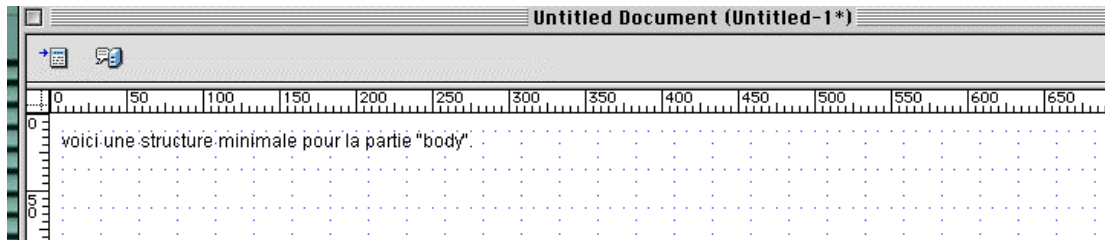
La partie "corps"

Dans le corps, on trouve le contenu à proprement parler d'une page html. Si la balise <body> reste vide, un navigateur (lecteur) ne visualisera rien du tout - les informations dans la partie "head" restent invisibles. Elles ne sont accessibles que si on ouvre la **page dite source** d'une page web . Voici une structure minimale de la partie <body> :

```
<html>
<head>
<title>Untitled Document</title>
<meta http-equiv="Content-Type" content="text/html; charset=iso-8859-1">
</head>

<body bgcolor="#FFFFFF">
<p>voici une structure minimale pour la partie &quot;body&quot;.</p>
</body>
</html>
```

Si on visualise cette instruction, on a la page suivante :



Voici une structure déjà bien plus élaborée du même contenu - de la même phrase - mise en valeur ... :

```

<html>
<head>
<title>page de test</title>
<meta http-equiv="Content-Type" content="text/html; charset=iso-8859-1">
</head>

<body bgcolor="#CCCCCC">
<p align="center">&nbsp;</p>
<p align="center">&nbsp;</p>
<p align="center">&nbsp;</p>
<p align="center">&nbsp;</p>
<table width="300" border="0" cellspacing="0" cellpadding="2" align="center" bgcolor="#CCFFFF"
bordercolordark="#FFFFFF" bordercolorlight="#333333" bordercolor="#FFFFFF">
  <tr>
    <td>
      <div align="center">
        <p>&nbsp;</p>
        <p><font color="#660033" face="Arial, Helvetica, sans-serif" size="5">voici
          une structure qui n'est plus minimale pour la partie &quot;body&quot;</font></p>
        <p>&nbsp;</p>
      </div>
    </td>
  </tr>
</table>
<p align="center">&nbsp;</p>
<p align="center">&nbsp;</p>
</body>
</html>

```

Dans la page suivante, on trouve la visualisation de la page correspondant à cette structure de balise. Afin d'éditer et de lire correctement une telle mise en forme d'une simple phrase, les instructions html se compliquent assez considérablement bien qu'on y trouve que 6 types de balises différentes :

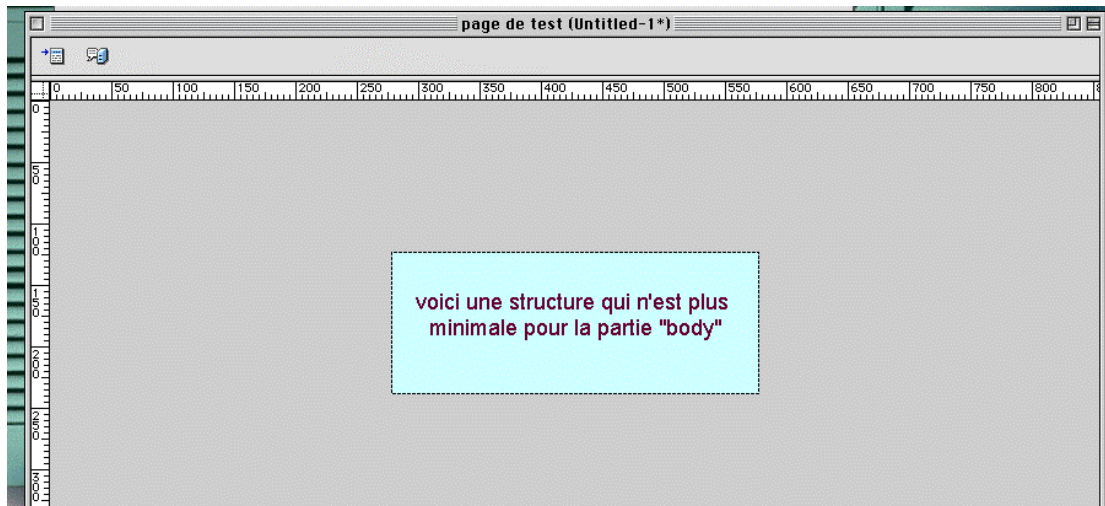
```

<body> : balise "contenu de la page"
<p> : balise "paragraphe"
<table> : balise tableau
<tr> : balise de ligne dans un tableau
<td> : balise de cellule dans un tableau
<div> : balise de "partie" (= division) du contenu

```

Mais, comme on le constatera, chacune de ces balises comporte un ensemble d'attributs nécessaires pour définir la page ci-dessus.

Voici donc la visualisation de la structure html introduite dans la page précédente.



Il faut noter que c'est la balise `<body>` elle-même qui peut être spécifiée par une diversité assez impressionnante d'attributs - attributs tels que :

- **bgcolor** : définit la couleur du fond de l'écran
- **text** : définit la couleur du texte
- **link** : définit la couleur des liens
- **vlink** : définit la couleur des liens déjà visités
- **alink** : définit la couleur des liens activés, c'est à dire quand le pointeur de la souris passe dessus
- **background** : définit l'image (gif ou jpeg) à utiliser comme fond d'écran
- **leftmargin** : Définit la largeur de la marge de gauche en pixels
- **topmargin** : Définit la largeur de la marge du haut en pixels

4) les commentaires

Enfin, pour rendre plus compréhensible une structure de balises ou encore pour signaler la propriété de la page, l'auteur de celle-ci, etc. on peut introduire, aussi bien dans la tête que dans le corps, voire même avant la déclaration <html>, des commentaires qui n'altèrent pas la dite structure. Exemple :

```
<!-- Ceci est un commentaire. -->
```

5) Références

site :

Publication HTML :

<http://www.ccim.be/ccim328/>

cours HTML

<http://nephi.unice.fr/CoursHTML/coursp1.html>

(encore à compléter)