De SGML à HTML : une histoire de balises

par Dominique Lahary
Bibliothèque départementale du Val-d'Oise

SGML[1]

Principe

SGML[2] permet de codifier la structure logique d'un document, son découpage en éléments constituant qui, le cas échéant, s'imbriquent les uns dans les autres. Cela permet à partir d'un même stockage électronique, de décliner une gamme de produits sur divers supports : papier, cédérom, accès en ligne.

Grâce à des normes associées, SGML est un outil d'échange. Il est également en train de devenir un outil de gestion de base de données car il permet de construire des accès sur des points donnés d'un document structuré et d'en gérer l'affichage d'éléments donnés.

SGML permet enfin de créer des liens internes (comme les appels de notes) et des liens entre documents, ce qui le rend apte à gérer le multimédia ainsi que l'hypertexte et l'hypermédia. Ainsi, Hy-Time, extension hypermédia de SGML, permet de gérer la synchronisation temporelle entre documents (y compris le son et l'image animée et sonorisée).

Comment fonctionne SGML ?

SGML repose sur un système de balisage des données selon une structure arborescente. Chaque élément logique du document est désigné par une balise qui est un code dont on décide qu'elle sera présente (-) ou omise (**O**) en début et/ou fin d'élément. Chaque élément peut se subdiviser en d'autres éléments eux-mêmes subdivisables, etc. La composition d'un élément en différents éléments de niveau immédiatement inférieur dans l'arborescence s'exprime à l'aide :

- de **connecteurs** :

, (les éléments doivent figurer dans l'ordre indiqué)
& (l'ordre des éléments est quelconque)
| (un seul élément doit figurer parmi ceux cités entre deux barres verticales)

- d'**indicateurs d'occurrence** :

? (l'élément est facultatif)
* (l'élément est facultatif et répétable)
+ (l'élément est obligatoire et répétable)
(par défaut, un élément est obligatoire et non répétable)

Voici un exemple fictif de structure HTML :

```
<!ELEMENT chapitre      - -  (titre, auteur+,
                              préambule?,section+)
<!ELEMENT titre         O O  (#PCDATA)
<!ELEMENT préambule     - O
<!ELEMENT recette       - O  (titrecet, ingrédients,
                              étapes+)
                                                [...]
```

1. Cette première partie relative à SGML doit beaucoup à Catherine Lupovici qui vint présenter SGML à l'Observatoire de l'information bibliographique de l'ABF le 28 mars 1995.
2. SGML : Standard Generalized Mark-up Language (langage normalisé et généralisé de balisage).

où un chapitre comprend un titre obligatoire et unique, un préambule unique et facultatif, des recettes obligatoires et multiples, comprenant un *titre* et une partie *ingrédients* obligatoires et non répétables et des paragraphes obligatoires éventuellement multiples, etc.

L'exemple cité, qui ne comprend pas de données mais ne fait qu'indiquer une structure, est une DTD (définition de type de document ou *document type definition*). Chaque DTD définit une structure propre à un type de document. Il existe des DTD qui sont des normes, ou des standards reconnus, d'autres propres à un utilisateur.

Cette DTD permet de représenter un document suivant le principe suivant :

```
<chapitre>
<titre>Entrées</titre>
<préambule>Les entrées sont dans la cuisine chinoise extrêmement variées...</préambule>
<recette>
<titrerecet>Crevettes frites au gingembre</titrerecet>
<paragraphe>300 g. de crevettes roses non décortiquées </paragraphe>
<paragraphe>1 petit morceau de gingembre frais </paragraphe>
[...]
<étape>Coupez les têtes et les pattes des crevettes puis essuyez-les soigneusement...</étape>
<étape>Faites chauffer de l'huile dans un wok...</étape>
[...]
</recette>
[...]
</chapitre>
```

On voit que la reconnaissance de la structure repose sur des balises ou marqueurs. La plupart des balises fonctionnent par couple : une balise de début, une balise de fin.

La balise <chapitre> proclame : « attention, un chapitre commence ». Et on demeure dans celui-ci tant que la balise </chapitre> n'a pas proclamé : « attention, ce chapitre est terminé ».

On voit également que la structure est arborescente : un livre comprend des chapitres qui comprennent des sections qui comprennent des paragraphes, etc. Le système des balises de début et de fin permet d'emboîter aisément ces composants.

SGML et les normes

L'ISO, qui avait reconnu SGML comme norme (ISO 8879), a adopté en 1994 les trois DTD de l'AAP[3] concernant respectivement les livres, les publications et séries et les articles de périodiques (ISO 12083). Cette norme est traduite en français[4], mais les codes de balise y demeurent en anglais.

HTML

HTML[5] est le format d'écriture des pages Web. Originellement, une telle page ne comprend que des indications de structure, susceptibles d'être interprétées différemment selon les navigateurs : c'est le client qui gère l'apparence physique du document. Mais on assiste à une évolution tendant à rapprocher les fonctionnalités de HTML de celles des traitements de texte, et les indications de mise en forme sont de plus en plus prégnantes : les créateurs d'information veulent maîtriser son apparence.

Une page HTML est un fichier. Il est d'usage de le nommer en lui donnant comme extension *htm* ou *html*. Dans l'exemple que nous développerons ci-dessous, nous écrirons le fichier *crevette.htm*.

3. AAP : American Association of Publishers.

4. Disponible à l'AFNOR : NF Z 71-010 (décembre 1990), 690 F.

5. HTML : Hyper Text Mark-up Language (langage de balisage hypertexte).

Structure générale

HTML utilise le principe de balisage de SGML et ses différentes versions constituent autant de DTD.

Une page HTML commence normalement par la balise <HTML> et se termine par </HTML>

Elle est composée d'un en-tête (*HEAD*) et un corps (*BODY*). L'en-tête comprend au moins un titre (*TITLE*), qui sera affiché sur l'écran de l'interrogateur en haut de la fenêtre de son navigateur. Quant au corps, il comprend la page proprement dite et d'éventuelles données de traitement.

La structure générale est donc la suivante :

```
<HTML>
<TITLE>Crevettes frites au
gingembre</TITLE>
<BODY>
[Texte de la page]
</BODY>
</HTML>
```

Les balises de mise en forme

Un certain nombre de balises permettent de donner une certaine forme au texte.

On peut par exemple déterminer des titres de tailles différentes. C'est l'objet des balises <H1> (la taille la plus grande) à H6 (la plus petite).

Exemple :

```
<H2>Crevettes frites au gingembre</H2>
<H4>Ingrédients<H4>
```

Comment aller à la ligne ? En utilisant au choix les balises <P> et </P> (paragraphe suivi d'une ligne inoccupée),
 (saut de ligne), <HR> (saut de ligne et trait horizontal).

Vous voulez faire apparaître du texte en italique ? Utilisez les balises <I> et </I>. En gras ? <B> et </B> :

```
300 g. de <B>crevettes roses</B> décortiquées
```

Pour centrer un texte, il suffit d'utiliser <CENTER> et </CENTER> :

```
<CENTER><H2>Crevettes frites
au gingembre</H2></CENTER>
```

On peut également dresser des listes : les balises <UL> et </UL> encadrent celle-ci, et chaque élément est introduit par <LI> :

```
<UL>
<LI>300 g. de crevettes roses décortiquées
<LI>1 petit morceau de gingembre
[...]
</UL>
```

Suivant le même principe, on peut établir des tableaux à plusieurs lignes et colonnes.

Les liens

Tout élément de texte peut être utilisé pour créer un lien vers un autre fichier ou un autre endroit du même fichier.
On utilise pour cela les balises <A HREF="xxx">yyy</A>, où *xxx* est l'adresse du fichier vers lequel on établit le lien et *yyy* l'élément, dit ancre, à partir duquel le lien est établi. S'agissant de texte, l'ancre apparaît généralement soulignée et l'utilisateur n'a qu'à cliquer pour accéder au fichier lié.

Pour établir un lien vers un autre site, il faut en faire figurer l'adresse entière :

```
<A HREF=
"http://www.gourmet.fr/pays.chine.html">
Autres sites sur la cuisine chinoise</A>
```

Mais si le fichier cible est sur le même site, on se contentera d'une adresse dite relative :

```
<A HREF= "sommaire.html">Sommaire</A>
```

Dans le premier exemple, *http* est le protocole propre au World Wide Web. Mais on peut établir un lien avec un autre protocole, par exemple *Telnet*, pour interroger directement un catalogue de bibliothèque qui ne dispose pas d'interface appropriée.

Pour insérer une image dans une page, on établit un lien avec le fichier la contenant, par la balise <IMG> contenant le nom de ce fichier :

```
<IMG SRC= "baguettes.jpg">
```

On peut faire de cette image l'ancre de départ d'un lien. Il suffit d'encadrer l'expression par les balises de lien :

```
<A HREF= "sommaire.htm"><IMG SRC=
"baguettes.jpg"></A>
```

L'image peut être fixe, mais aussi animée (une séquence vidéo) et on peut également insérer du son.

Les entités

HTML ne gère en principe qu'un jeu limité de caractères, et les lettres accentuées doivent être remplacées par des entités codées. C'est ainsi que :

é est représenté par *é*

ê par *ê*

à par *à*

Mais certains navigateurs interprètent correctement les caractères accentués sans qu'il soit besoin de les remplacer par ces codes.

Exemple

Nous en savons maintenant assez pour écrire la page suivante :

```
<HTML>
<TITLE>Crevettes frites au gingembre</TITLE>
<BODY>
<CENTER>La cuisine chinoise : Les entr&eacute;es</CENTER>
<CENTER><H2>Crevettes frites au gingembre</H2>
Recette pour 4 personnes<BR>
Pr&eacute;paration et cuisson 20 minutes</CENTER>
<H4>Ingr&eacute;dients</H4>
<UL>
<LI>300 g. de <B>crevettes roses</B> d&eacute;cortiqu&eacute;es
<LI>1 petit morceau de <B>gingembre</B>
<LI>1 petit morceau d'<B>oignon nouveau</B>
<LI>2 cuiller&eacute;es &agrave; soupe de <B>sauce de soja</B>
<LI>2 cuiller&eacute;es &agrave; caf&eacute; de <B>sucre</B>
<LI>2 cuiller&eacute;es &agrave; soupe de <B>vin de Shaoxing</B>
<LI>1/2 cuiller&eacute;e &agrave; caf&eacute; de <B>sel</B>
<LI><B>De l'<B>huile</B> pour friture</B>
<LI> 1 cuiller&eacute;e &agrave; soupe de <B>coriandre</B> hach&eacute;e.
</UL>
<HR><P>Coupez les t&ecirc;tes et les pattes des crevettes. Pelez et hachez le gingembre et l'oignon.</P>
<P>Faites chauffer de l'huile dans un wok et plongez-y les crevettes. Faites-les frire jusqu'&agrave; ce qu'elles deviennent rouges, puis retirez-les rapidement.</P>
<P>Retirez l'huile du wok et mettez-y le gingembre, l'oignon, le sucre, la sauce de soja et le vin de Shaoxing. Ajoutez les crevettes et m&eacute;langez. Disposez sur un plat et parsemez-le de coriandre.</P>
<P>Servez chaud ou froid.</P>
<CENTER><A HREF= "sommaire.html"><IMG SRC= "baguettes.jpg"></A></CENTER>
<CENTER><A HREF= "sommaire.html">Sommaire</A> – <A HREF= "entrees.html">Entr&eacute;es
</A> – <A HREF= "index.html">Index</A> –
<A HREF= "http://www.gourmet.fr/pays.chine.html">Autres sites sur la cuisine chinoise</A></CENTER>
</BODY>
</HTML>
```

Cette page pourra être ainsi interprétée par le navigateur :

La cuisine chinoise : Les entrées

Crevettes frites au gingembre

Recette pour 4 personnes

Préparation et cuisson 20 minutes

Ingrédients

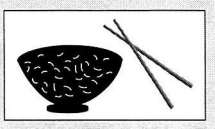
- 300 g. de crevettes roses décortiquées
- 1 petit morceau de gingembre
- 1 petit morceau d'oignon nouveau
- 2 cuillerées à soupe de sauce de soja
- 2 cuillerées à café de sucre
- 2 cuillerées à soupe de vin de Shaoxing
- 1/2 cuillerée à café de sel
- De l'huile pour friture
- 1 cuillerée à soupe de coriandre hachée.

Coupez les têtes et les pattes des crevettes. Pelez et hachez le gingembre et l'oignon.

Faites chauffer de l'huile dans un wok et plongez-y les crevettes. Faites-les frire jusqu'à ce qu'elles deviennent rouges, puis retirez-les rapidement.

Retirez l'huile du wok et mettez-y le gingembre, l'oignon, le sucre, la sauce de soja et le vin de Shaoxing. Ajoutez les crevettes et mélangez. Disposez sur un plat et parsemez-le de coriandre.

Servez chaud ou froid.



Sommaire – Entrées – Index – Autres sites sur la cuisine chinoise

HTML et les bases de données

On rencontre de plus en plus de bases de données directement interrogeables sur le World Wide Web. Pour qu'une base soit ainsi accessible, il faut qu'une interface, dite CGI (*common gateway interface*), ait été programmée.

Le site adresse aux utilisateurs une page HTML de type formulaire, qui comprend des cases dans lesquelles une requête peut être saisie, et éventuellement des cases à cocher ou des sélections à effectuer en cliquant sur un menu déroulant.

Le serveur reçoit une question, la transmet à la base de données et traduit sa réponse en page HTML. Celle-ci n'est pas un fichier statique, stockée tel quel sur le serveur : elle a été générée à l'occasion de la requête. Elle peut, le cas échéant, contenir des images, du son et des ancres créant des liens vers des documents, mais aussi relançant une requête.

Java

Dernier développement, les pages HTML peuvent contenir de petits programmes appelés *applets* que le navigateur, s'il sait les interpréter, va exécuter sur le poste client. Ceci grâce au langage Java, lancé par Sun en 1995, ou une variante simplifiée Javascript.

Ces *applets* permettent d'animer les pages, en faisant défiler du texte, en déplaçant ou modifiant une image, mais aussi d'opérer des traitements comme la conversion de devises ou la génération de calendriers. Au-delà de ces exemples dont certains peuvent paraître futiles, Java apparaît comme un langage indépendant des matériels et systèmes d'exploitation souvent présenté comme extrêmement prometteur.

HTML est-il un standard ?

Inventé par Tim Berners Lee au CERN de Genève, HTML est géré par le World Wide Web Consortium (W3C), organisme international comprenant un certain nombre d'acteurs du World Wide Web, dont des entreprises privées. Le W3C est responsable de l'évolution du langage et en publie les versions successives. HTML 2.0 est encore, au début de 1997, la version la plus répandue, et la dernière en date est la version HTML 3.2.

Parallèlement à cette évolution officielle, quoi que ne relevant d'aucun organisme de normalisation proprement dit, HTML fait l'objet d'aménagements privés, appelés *extensions*, notamment des deux principaux fournisseurs de navigateur : Netscape pour Netscape Navigator et Microsoft pour Internet Explorer. Certaines extensions sont correctement interprétées par plusieurs navigateurs, d'autres par un seul. On ne peut que souhaiter le respect d'un standard commun, par les éditeurs de navigateurs comme par ceux de sites Web.