

Fiche XML n°4

Encodage et caractères en XML

Cours

1. Codage des caractères

XML est un standard du W3C permettant de structurer des documents contenant des textes et des données.

Souvent, les textes sont dépendants d'une langue et donc de caractères spécifiques (chinois, japonais, finnois, danois, arabe, hébreux, grec, anglais, français...). Or, un standard est par essence même une structure commune et reconnue par tous. Un fichier XML doit donc être lisible¹ quelque soit le lieu et l'outil ou la personne qui le consulte.

Y-a-t-il contradiction ? Non, le W3C et les organismes de normalisation ont prévu un jeu de caractères (presque) universel a été mis au point à cet effet : Unicode².

2. Unicode

Unicode est un jeu de caractères standard et international. La version 3.2 de mai 2002 contenait déjà 95156 caractères ! Il couvre tous les symboles des différentes langues les plus connues :

- anglais, français (alphabet Latin en général) ;
- grec ancien et moderne, écritures cyrilliques ;
- han (Japon et Chine), hangûl (Corée) et leur représentations phonétiques katana, hiragana ;
- arabe et hébreu ;
- devanagari, thaï, bengali, tibétain (Sous-continent Indien)...

Au total, moins d'une personne sur mille parle une langue qui n'est pas représentée par Unicode. Du point de vue informatique, le type de codage associé est intitulé UTF-8 (mais aussi UTF-16).

3. Autres jeux de caractères

Unicode est toutefois très récent. D'autres codes plus «régionaux» et plus anciens existent. Ils sont essentiellement des extensions du code ASCII³.

Les plus courants sont les normes de l'ISO⁴ :

- ISO-8859-1 (latin-1) : allemand, danois, espagnol, français, finnois, hollandais, italien, islandais, norvégien, portugais et suédois ;
- ISO-8859-2 (latin-2) : allemand, anglais, croate, hongrois, polonais, roumain, slovaque, slovène, serbe et tchèque ;
- ISO-8859-3 (latin-3) : espéranto, maltais et turc ;
- ISO-8859-6 : arabe de base ;
- ISO-8859-8 : écriture hébraïque (hébreu et yiddish) ;
- ISO-8859-4, ISO-8859-5, ISO-8859-7, ISO-8859-9... ISO-8859-16.

Il existe d'autres standard régionaux, comme le KS C 5601-1992 (coréen), ou liés à des environnements informatiques comme Cp1252 (Microsoft ; proche du ISO-8859-1) ou MacRoman (Apple).

¹ En revanche, il n'est pas forcément compréhensible si la personne ne connaît pas la langue utilisée !

² <http://unicode.org/>

³ La norme ASCII (American Standard Code for Information Interchange) est la norme de codage de caractères en informatique la plus connue et la plus largement compatible. (<http://fr.wikipedia.org/wiki/ASCII>)

⁴ L'International Organization for Standardization, ou ISO, est un organisme de normalisation international composé de représentants d'organisations nationales de normalisation de différents pays. (<http://fr.wikipedia.org/wiki/ISO>)

4. Définir et utiliser un jeu de caractères en XML

Entité XML	Symbole
&	&
<	<
>	>
"	"
'	'

Unicode est le jeu par défaut en XML. Donc, si rien n'est indiqué, le document est supposé être encodé en Unicode (utf-8). Cependant, il est possible de préciser l'encodage en déclarant l'utilisation d'un jeu de caractères en ajoutant l'attribut «`encoding="jeu"`» dans la ligne d'entête XML. Ainsi, «`<?xml version="1.0" encoding="ISO-8859-1" ?>`» indique que le document XML est en latin-1, norme associée au français.

Une difficulté dans certains cas est d'accéder aux symboles. Pour presque tous les symboles ASCII étendu, n'importe quel clavier et éditeur de texte suffisent. Certains caractères ne sont pas utilisables directement : ceux réservés aux balises, aux attributs et aux entités. Ils doivent être codés pour être utilisés en dehors des balises. Le tableaux ci-contre présente ces quelques caractères.

Pour les caractères qui ne sont pas ASCII étendus, il faut y accéder par leur codage (UTF-8 par exemple) en décimal, en hexadécimal ou, pour certains appartenants au jeu d'entités XHTML, par leur identifiant. Des tables sont à disposition pour accéder à leur code. Soit «`C`» le code d'un caractère, son utilisation dans le document sera «`&#C;`» si «`C`» est décimal et «``» s'il est hexadécimal. Par exemple, pour le symbole 'À' on utilisera : «`À`» ou «`À`».

5. Langue d'un document XML

Unicode prend en compte de nombreuses langues. Il n'est pas facile d'identifier (automatiquement) la langue liée à un texte. Pour cela, il existe un attribut XML spécifique «`xml:lang`». C'est le seul nom XML réservé à XML dans la version 1.0. La valeur de cet attribut est prise dans la norme internationale ISO-639⁵. L'exemple suivant illustre un document XML bien formé écrit en français à l'aide d'Unicode.

6. Entités générales et Unicode

Les entités générales peuvent être utilisées pour définir des symboles spécifiques liés aux caractères Unicode. Comme nous l'avons déjà vu, en XHTML, il existe déjà des entités associées à des symboles.

```
<?xml version='1.0' encoding='UTF-8' ?>
<racine xml:lang='fr'>Le serveur principal se trouve dans la salle
s&#xE9;curis&#xE9;e. Il g&#xE8;re le service de courrier &#xE9;lectronique et
le serveur web (Apache)... </racine>
```

Pour les charger, il faut procéder en utilisant des entités paramètres.

Par exemple, «À» peut s'écrire aussi : «`À`» si la DTD XHTML définissant cette entité est référencée (par exemple par l'URL «<http://www.w3.org/TR/xhtml1/DTD/xhtml-lat1.ent>»). Dans l'exemple ci-dessous, «`⇒`» est le symbole «`⇒`» en XHTML et est introduit par «`%SymbolesHTML;`».

```
<?xml version='1.0' encoding='UTF-8' ?>
<!DOCTYPE racine [
  <!ENTITY euro '&#x20AC;'>
  <!ENTITY % SymbolesHTML PUBLIC '-//W3C//ENTITIES Symbols for XHTML//EN'
    'http://www.w3.org/TR/xhtml1/DTD/xhtml-symbol.ent' >
  %SymbolesHTML;]>
<racine>Le produit vaut 20 &euro; &rArr; c'est cher !</racine>
```

Aux critères permettant de déterminer si un document est un document XML bien formé, il faut donc ajouter la règle suivante : il faut obligatoirement utiliser l'entité associée à certains caractères comme '<', '&'... dans les informations textuelles.



Cette création est mise à disposition selon le Contrat Paternité-Pas d'Utilisation Commerciale-Pas de Modification 2.0 France disponible en ligne <http://creativecommons.org/licenses/by-nc-nd/2.0/fr/> ou par courrier postal à Creative Commons, 171 Second Street, Suite 300, San Francisco, California 94105, USA.

⁵ L'ISO 639 (ICS n°01.140.20) est une norme internationale l'ISO qui définit des codes pour la représentation des noms de langues. (http://fr.wikipedia.org/wiki/ISO_639)