

# R109 : Séance de TP 1

---

## Sommaire

1. [HTML vs XHTML](#)
  2. [Codage UTF-8](#)
    1. [Jeu universel de caractères codés](#)
    2. [Unicode](#)
    3. [UTF-8](#)
  3. [Autres codages](#)
    1. [ISO-8859-1](#)
    2. [Windows 1252](#)
- 

## HTML vs XHTML

HTML est langage de balises issu du langage SGML. XHTML est une refonte de HTML à partir du langage XML, plus strict que SGML. Comme XHTML et HTML n'ont pas les mêmes langages parents, des différences existent :

- en XHTML les *noms des balises* et les *noms des attributs* doivent obligatoirement être écrits en *minuscules*
- les *valeurs des attributs* doivent obligatoirement être écrits entre *guillemets*
- toute balise ouverte doit être fermée, en particulier, les balises vides du HTML deviennent des balises autofermantes par ajout d'une barre oblique avant le chevron fermant de la balise, par exemple `<br>` devient `<br/>`
- tous les éléments doivent être explicitement balisés, en particulier, l'en-tête et le corps d'un document doivent être explicitement balisés par `<head>...</head>` et `<body>...</body>`
- tous les attributs doivent avoir une valeur explicite (pas d'attributs condensés)

## Codage UTF-8

### Jeu universel de caractères codés

La norme [ISO/CEI-10646](#) définit une table de caractères appelée « jeu universel de caractères codés » (JUC), ou « universal character set » (UCS) en anglais. Cette table contient environ 110000 caractères issus du monde entier. Chaque caractère du JUC est identifié par un point de code noté U+x, où x **est un entier hexadécimal** positif comportant de 4 à 6 chiffres :

- 4 chiffres pour le plan multilingue de base (PMB), de U+0000 à U+FFFF
- 5 chiffres pour les 15 plans suivants, de U+10000 à U+FFFFF
- 6 chiffres pour le dernier plan, de U+100000 à U+10FFFF

Chaque caractère du JUC possède également un nom, par exemple, la lettre A majuscule de notre alphabet latin s'appelle « LATIN CAPITAL LETTER A » et a pour point de code U+0041.

### Unicode

Le standard [Unicode](#) est lié à la norme [ISO/CEI-10646-10646](#) car il reprend le JUC et y ajoute des règles de gestion : collation, codage, sérialisation... [Unicode](#) accepte plusieurs méthodes de codage pour représenter un point de code valide : UTF-8, UTF-16 et UTF-32. Le nombre derrière le sigle UTF est le nombre de bits minimal d'une unité de code, appelée « codet ».

### UTF-8

Le codage le plus utilisé est [UTF-8](#), qui est un codage de longueur variable dont les codets sont des unités de 8 bits.

En [UTF-8](#) un point de code U+n se code sur un, deux, trois ou quatre octets en fonction du nombre de bits nécessaires pour coder n. Si n appartient au code [ASCII](#) (7 bits), alors n est compris entre 00 et 7F et se code sur un seul octet avec 0 comme bit de poids fort.

[UTF-8](#) est donc entièrement compatible avec le code [ASCII](#).

Dans les autres cas, les bits de poids fort du premier octet commencent par une suite de 1 suivie d'un 0. **Le nombre de 1 indique le nombre total d'octets dans le codage de U+n**. Les deux bits de poids fort de tous les octets qui suivent portent toujours la marque 10. Un point de code codé en [UTF-8](#) aura donc l'une des quatre formes suivantes :

1. 0xxxxxxx : jusqu'à 7 bits
2. 110xxxxx 10xxxxxx : entre 8 et 11 bits
3. 1110xxxx 10xxxxxx 10xxxxxx : entre 12 et 16 bits
4. 11110xxx 10xxxxxx 10xxxxxx 10xxxxxx : entre 17 et 21 bits

Quand plusieurs formes sont possibles, *seule la plus courte est considérée comme valide*.

Les bits x correspondent simplement au codage de n en binaire cadré à droite et éventuellement complété par des 0 à gauche.

# Autres codages

## ISO-8859-1

| ISO-8859-1 |      |     |     |     |     |     |     |     |     |      |     |     |     |     |     |     |
|------------|------|-----|-----|-----|-----|-----|-----|-----|-----|------|-----|-----|-----|-----|-----|-----|
|            | x0   | x1  | x2  | x3  | x4  | x5  | x6  | x7  | x8  | x9   | xA  | xB  | xC  | xD  | xE  | xF  |
| 0x         | NUL  | SOH | STX | ETX | EOT | ENQ | ACK | BEL | BS  | HT   | LF  | VT  | FF  | CR  | SO  | SI  |
| 1x         | DLE  | DC1 | DC2 | DC3 | DC4 | NAK | SYN | ETB | CAN | EM   | SUB | ESC | FS  | GS  | RS  | US  |
| 2x         | SP   | !   | "   | #   | \$  | %   | &   | '   | (   | )    | *   | +   | ,   | -   | .   | /   |
| 3x         | 0    | 1   | 2   | 3   | 4   | 5   | 6   | 7   | 8   | 9    | :   | ;   | <   | =   | >   | ?   |
| 4x         | @    | A   | B   | C   | D   | E   | F   | G   | H   | I    | J   | K   | L   | M   | N   | O   |
| 5x         | P    | Q   | R   | S   | T   | U   | V   | W   | X   | Y    | Z   | [   | \   | ]   | ^   | _   |
| 6x         | `    | a   | b   | c   | d   | e   | f   | g   | h   | i    | j   | k   | l   | m   | n   | o   |
| 7x         | p    | q   | r   | s   | t   | u   | v   | w   | x   | y    | z   | {   |     | }   | ~   | DEL |
| 8x         | PAD  | HOP | BPH | NBH | IND | NEL | SSA | ESA | HTS | HTJ  | VTS | PLD | PLU | RI  | SS2 | SS3 |
| 9x         | DCS  | PU1 | PU2 | STS | CCH | MW  | SPA | EPA | SOS | SGCI | SCI | CSI | ST  | OSC | PM  | APC |
| Ax         | NBSP | ı   | ç   | £   | ¤   | ¥   | ¦   | §   | ¨   | ©    | ª   | «   | ¬   | ®   | ¯   |     |
| Bx         | °    | ±   | ²   | ³   | ´   | µ   | ¶   | ·   | ¸   | ¹    | º   | »   | ¼   | ½   | ¾   | ¿   |
| Cx         | À    | Á   | Â   | Ã   | Ä   | Å   | Æ   | Ç   | È   | É    | Ê   | Ë   | Ì   | Í   | Î   | Ï   |
| Dx         | Ð    | Ñ   | Ò   | Ó   | Ô   | Õ   | Ö   | ×   | Ø   | Ù    | Ú   | Û   | Ü   | Ý   | Þ   | ß   |
| Ex         | à    | á   | â   | ã   | ä   | å   | æ   | ç   | è   | é    | ê   | ë   | ì   | í   | î   | ï   |
| Fx         | ð    | ñ   | ò   | ó   | ô   | õ   | ö   | ÷   | ø   | ù    | ú   | û   | ü   | ý   | þ   | ÿ   |

page de codes cp859

## Windows-1252

| Windows-1252 (CP1252) |      |     |     |     |     |     |     |     |     |    |     |     |    |    |    |     |
|-----------------------|------|-----|-----|-----|-----|-----|-----|-----|-----|----|-----|-----|----|----|----|-----|
|                       | x0   | x1  | x2  | x3  | x4  | x5  | x6  | x7  | x8  | x9 | xA  | xB  | xC | xD | xE | xF  |
| 0x                    | NUL  | SOH | STX | ETX | EOT | ENQ | ACK | BEL | BS  | HT | LF  | VT  | FF | CR | SO | SI  |
| 1x                    | DLE  | DC1 | DC2 | DC3 | DC4 | NAK | SYN | ETB | CAN | EM | SUB | ESC | FS | GS | RS | US  |
| 2x                    | SP   | !   | "   | #   | \$  | %   | &   | '   | (   | )  | *   | +   | ,  | -  | .  | /   |
| 3x                    | 0    | 1   | 2   | 3   | 4   | 5   | 6   | 7   | 8   | 9  | :   | ;   | <  | =  | >  | ?   |
| 4x                    | @    | A   | B   | C   | D   | E   | F   | G   | H   | I  | J   | K   | L  | M  | N  | O   |
| 5x                    | P    | Q   | R   | S   | T   | U   | V   | W   | X   | Y  | Z   | [   | \  | ]  | ^  | _   |
| 6x                    | `    | a   | b   | c   | d   | e   | f   | g   | h   | i  | j   | k   | l  | m  | n  | o   |
| 7x                    | p    | q   | r   | s   | t   | u   | v   | w   | x   | y  | z   | {   |    | }  | ~  | DEL |
| 8x                    | €    | ‚   | ƒ   | „   | …   | †   | ‡   | ^   | ‰   | Š  | ‹   | Œ   | Ž  |    |    |     |
| 9x                    | ‘    | ’   | “   | ”   | •   | –   | —   | ~   | ™   | š  | ›   | œ   | ž  | ÿ  |    |     |
| Ax                    | NBSP | ı   | ç   | £   | ¤   | ¥   | ¦   | §   | ¨   | ©  | ª   | «   | ¬  | ®  | ¯  |     |
| Bx                    | °    | ±   | ²   | ³   | ´   | µ   | ¶   | ·   | ¸   | ¹  | º   | »   | ¼  | ½  | ¾  | ¿   |
| Cx                    | À    | Á   | Â   | Ã   | Ä   | Å   | Æ   | Ç   | È   | É  | Ê   | Ë   | Ì  | Í  | Î  | Ï   |
| Dx                    | Ð    | Ñ   | Ò   | Ó   | Ô   | Õ   | Ö   | ×   | Ø   | Ù  | Ú   | Û   | Ü  | Ý  | Þ  | ß   |
| Ex                    | à    | á   | â   | ã   | ä   | å   | æ   | ç   | è   | é  | ê   | ë   | ì  | í  | î  | ï   |
| Fx                    | ð    | ñ   | ò   | ó   | ô   | õ   | ö   | ÷   | ø   | ù  | ú   | û   | ü  | ý  | þ  | ÿ   |

page de codes cp1252

