# Caractères français et ordre alphabétique

Jacques.Andre@irisa.fr

3 mai 2003

Ce document est essentiellement formé d'un tableau des lettres françaises triées par ordre alphabétique avec pour chacune leur code en Ascii (qaund il existe), en Latin-1 ou 9, en Unicode et la façon de les coder en TEX, Mime et HTML. Par ailleurs on montre comment lire les courriers électroniques qui ont été codés en UTF-8 et lus par un navigateur qui travaille en Latin-1, par exemple passer de « La rÃ©fÃ©rence franÃ§aise considÃ¨re » à « La référence française considére » !

L'inventaire des lettres françaises et leur ordre alphabétique est issu de :

> Alain LABONTÉ, « Règles du classement alphabétique en langue française et procédure informatisée pour le tri », Conseil du trésor, Québec, Canada, 2002 : `http://www.tresor.gouv.qc.ca/doc/classm.htm`

document qui sert de base à son auteur pour rédiger un projet de norme internationale de classement (projet ISO/CEI 14651) pour l'ensemble des caractères du jeu universel de caractères codés sur plusieurs octets (norme ISO/CEI 10646-1:1993, correspondant au standard UNICODE).

## Autres liens utiles

- Liste typographie (où nombre de discussions ont eu lieu sur ce thème, notamment en février 2003 sous la plume de Jean-François Robert, Jean Fontaine et ... Alain laBonté) : `https://www.irisa.fr/wws/info/typographie`
- Les archives du forum `fr.lettres.langue.francaise` à `http://www.langue-fr.net/`.
- Le site `http://pages.infinit.net/hapax` avec la traduction française d'Unicode et une introduction de Patrick Andries à Unicode.
- Voir aussi `http://www.eki.ee/letter/` pour divers inventaires de caractères par langues et le transcodage d'un code à l'autre.

## Légende du tableau des lettres françaises

**Car.** le caractère (ou plutôt un glyphe au sens d'Unicode).

**Ascii** son code en Ascii, quand il existe (sinon il est marqué « — »), en base 10 (on retrouvera cette même valeur en octal en colonne « Latin » et en hexadécimal en colonne « Unicode ».

**TEX** son nommage en TEX, en n'utilisant que les caractères Ascii.

**HTML** le nommage des caractères par des entités (écrites en Ascii). On peut aussi les nommer sous la forme `&#x....;` où `....` en est le code Unicode (colonne Unicode), par exemple saisir `&#x0178;` pour avoir Ÿ.

**QP** son codage en Mime *quoted printable*[1] (comparer avec le code Unicode).

**Latin** son codage en ISO 8859-1 (Latin-1) et -15 (Latin-9). Lorsque le code diffère de Latin-1 à Latin-9 (c'est le seul cas des caractères œ, Œ et Ÿ), la première ligne indique « — » (pas de code en Latin-1) et la seconde le code en Latin-9.

**Unicode** son codage en Unicode, en hexadécimal.

**UTF-8** la forme de stockage par 8 bits du code Unicode : le principe de cette forme est de recoder sur 1 octet les caractères Unicode de code U+0000 à U+007F (c'est-à-dire ceux de l'Ascii), sur 2 octets les codes de U+0080 à U+07FF, etc.[2] Dans le second cas, le principe est de répartir ainsi les bits :

`0000 0yyy yyxx xxxx => 110y yyyy 10xx xxxx`

mais pour les caractères français de Latin-1 : `yyy yy=000 11`; en effet « À », le premier d'entre eux, a pour code U+00C0 soit `1100 0000`. Tous les codes UTF-8 des caractères français avec diacritique vont donc avoir pour premier octet la valeur `C3` (on donne ici leur code en hexa, et on sépare les deux octets éventuels par un espace) sauf donc ceux de œ, Œ et Ÿ pour lequel ce premier octet sera `C5`.

⇒ si on reçoit dans un courrier électronique un texte français stocké en UTF-8 et qu'on le lit comme si c'était du Latin-1 (c'est ce qui arrive souvent par défaut quand on n'a pas paramétré son navigateur pour recevoir de l'UTF-8) les lettres de l'Ascii étant codées sur 1 octet apparaissent bien ; les autres sont sur 2 octets interpétrés comme 2 lettres : la première, de code C3 (ou plus rarement C5) apparaît alors comme Ã (ou Å) ; la seconde, de code binaire `10xx xxxx`, va correspondre soit à un des caractères de commande de Latin-1 (codes 80 à 9F) qui ne sont pas imprimables[3] et n'apparaissent alors en général pas (nous les indiquons ici par un □), soit à un des premiers caractères spéciaux de Latin-1 (codes $A0_{16}$ à $BF_{16}$. Notons que parmi ces derniers se trouvent des caractères qui varient de Latin-1 à Latin-9, par exemple $A6_{16}$ (¦ en Latin-1 et Š en Latin-9).

---

[1] On retrouve ici ce qui apparaît dans le courrier électronique en mode *quoted printable*, codage sur 7 bits de caractères 8 bits. Voir `http://www.ietf.org/`.

[2] Voir la transformation détaillée dans `http://staff.dstc.edu.au/ilister/utf8.html`.

[3] Sauf lorsque des standards propriétaires réutilisent ces codes pour y mettre des caractères qui ne sont pas de Latin-1, c'est par exemple le cas des Mac, de certains codages de Windows, voire de... TEX.

## Tableau des lettres françaises triées par ordre alphabétique

| Car. base | Ascii 10 | TEX Ascii | HTML Ascii | QP Ascii | Latin 8 | Unicode 16 | UTF-8 16 | ⇒ lat1 |
|---|---|---|---|---|---|---|---|---|
| a | 97 | a | a | a | 141 | 0061 | 61 | a |
| A | 65 | A | A | A | 101 | 0041 | 41 | A |
| à | — | \'a | à | =E0 | 340 | 00E0 | C3 A0 | Ã |
| À | — | \`A | À | =C0 | 300 | 00C0 | C3 80 | Ã□ |
| â | — | \^a | â | =E2 | 342 | 00E2 | C3 A2 | Ã¢ |
| Â | — | \^a | Â | =C2 | 302 | 00C2 | C3 82 | Ã□ |
| æ | — | {\ae} | æ | =E6 | 346 | 00E6 | C3 A6 | Ã¦ |
| Æ | — | {\AE} | Æ | =C6 | 306 | 00E6 | C3 86 | Ã□ |
| b | 98 | b | b | b | 142 | 0062 | 62 | b |
| B | 66 | B | B | B | 102 | 0042 | 42 | B |
| c | 99 | c | c | c | 143 | 0063 | 63 | c |
| C | 67 | C | C | C | 103 | 0043 | 43 | C |
| ç | — | \c{c} | ç | =E7 | 347 | 00E7 | C3 A2 | Ã§ |
| Ç | — | \c{C} | Ç | =C7 | 307 | 00C7 | C3 87 | Ã□ |
| d | 100 | d | d | d | 144 | 0064 | 64 | d |
| D | 68 | D | D | D | 104 | 0044 | 44 | D |
| e | 101 | e | e | e | 145 | 0065 | 65 | e |
| E | 69 | E | E | E | 105 | 0045 | 45 | E |
| é | — | \'e | é | =E9 | 351 | 00E9 | C3 A9 | Ã© |
| É | — | \'E | É | =C9 | 311 | 00C9 | C3 89 | Ã□ |
| è | — | \`e | è | =E8 | 350 | 00E8 | C3 A8 | Ã¨ |
| È | — | \`E | È | =C8 | 310 | 00C8 | C3 88 | Ã□ |
| ê | — | \^e | ê | =EA | 352 | 00EA | C3 AA | Ãª |
| Ê | — | \^E | Ê | =CA | 312 | 00CA | C3 8A | Ã□ |
| ë | — | \"e | ë | =EB | 353 | 00EB | C3 AB | Ã« |
| Ë | — | \"E | Ë | =CB | 313 | 00CB | C3 8B | Ã□ |
| f | 102 | f | f | f | 146 | 0066 | 66 | f |
| F | 70 | F | F | F | 106 | 0046 | 46 | F |
| g | 103 | g | g | g | 147 | 0067 | 67 | g |
| G | 71 | G | G | G | 107 | 0047 | 47 | G |
| h | 104 | h | h | h | 150 | 0068 | 68 | h |
| H | 72 | H | H | H | 110 | 0048 | 48 | H |
| i | 105 | i | i | i | 151 | 0069 | 69 | i |
| I | 73 | I | I | I | 111 | 0049 | 49 | I |
| î | — | \^{\i} | î | =EE | 356 | 00EE | C3 AE | Ã® |
| Î | — | \^{\I} | Î | =CE | 316 | 00CE | C3 CE | Ã□ |
| ï | — | \"{\i} | ï | =EF | 357 | 00EF | C3 AF | Ã¯ |
| Ï | — | \"{\I} | Ï | =CF | 317 | 00CF | C3 CF | Ã□ |
| j | 106 | j | j | j | 152 | 006A | 6A | j |
| J | 74 | J | J | J | 112 | 004A | 4A | J |
| k | 107 | k | k | k | 153 | 006B | 6B | k |
| K | 75 | K | K | K | 113 | 004B | 4B | K |
| l | 108 | l | l | l | 154 | 006C | 6C | l |
| L | 76 | L | L | L | 114 | 004C | 4C | L |

| Car. base | Ascii 10 | TEX Ascii | HTML Ascii | QP Ascii | Latin 8 | Unicode 16 | UTF-8 16 | ⇒ lat1 |
|---|---|---|---|---|---|---|---|---|
| m | 109 | m | m | m | 155 | 006D | 6D | m |
| M | 77 | M | M | M | 115 | 004D | 4D | M |
| n | 110 | n | n | n | 156 | 006E | 6E | n |
| N | 78 | N | N | N | 116 | 004E | 4E | N |
| o | 111 | o | o | o | 157 | 006F | 6F | o |
| O | 79 | O | O | O | 117 | 004F | 4F | O |
| ô | — | \^o | ô | =F4 | 364 | 00F4 | C3 B4 | Ã´ |
| Ô | — | \^O | Ô | =D4 | 324 | 00D4 | C3 94 | Ã□ |
| œ | — | {\oe} | œ | — | — 275 | 0153 | C5 93 | Å□ |
| Œ | — | {\OE} | Œ | — | — 274 | 0152 | C5 92 | Å□ |
| p | 112 | p | p | p | 160 | 0070 | 70 | p |
| P | 80 | P | P | P | 120 | 0050 | 50 | P |
| q | 113 | q | q | q | 161 | 0071 | 71 | q |
| Q | 81 | Q | Q | Q | 121 | 0051 | 51 | Q |
| r | 114 | r | r | r | 162 | 0072 | 72 | r |
| R | 82 | R | R | R | 122 | 0052 | 52 | R |
| s | 115 | s | s | s | 163 | 0073 | 73 | s |
| S | 83 | S | S | S | 123 | 0053 | 53 | S |
| t | 116 | t | t | t | 164 | 0074 | 74 | t |
| T | 84 | T | T | T | 124 | 0054 | 54 | T |
| u | 117 | u | u | u | 165 | 0075 | 75 | u |
| U | 85 | U | U | U | 125 | 0055 | 55 | U |
| ù | — | \‘u | ù | =F9 | 371 | 00F9 | C3 B9 | Ã[1] |
| Ù | — | \’U | Ù | =D9 | 331 | 00D9 | C3 99 | Ã□ |
| û | — | \^u | û | =FB | 373 | 00FB | C3 BB | Ã» |
| Û | — | \^U | Û | =DB | 333 | 00DB | C3 9B | Ã□ |
| ü | — | \"u | ü | =FC | 374 | 00FC | C3 BC | Ã¼ |
| Ü | — | \"U | Ü | =DC | 334 | 00DC | C3 9C | Ã□ |
| v | 118 | v | v | v | 166 | 0076 | 76 | v |
| V | 86 | V | V | V | 126 | 0056 | 56 | V |
| w | 119 | w | w | w | 167 | 0077 | 77 | w |
| W | 87 | W | W | W | 127 | 0057 | 57 | W |
| x | 120 | x | x | x | 170 | 0078 | 78 | x |
| X | 88 | X | X | X | 130 | 0058 | 58 | X |
| y | 121 | y | y | y | 171 | 0079 | 79 | y |
| Y | 89 | Y | Y | Y | 131 | 0059 | 59 | Y |
| ÿ | — | \"y | ÿ | =FF | 377 | 00FF | C3 BF | Ã¿ |
| Ÿ | — | \"Y | Ÿ | — | — 276 | 0178 | C5 B8 | Ã¸ |
| z | 122 | z | z | z | 172 | 007A | 7A | z |
| Z | 90 | Z | Z | Z | 132 | 005A | 5A | Z |

Page accessible à `http://www.irisa.fr/faqtypo/unicode/alpha-fr.pdf`