Chapitre 15

Tableaux de codes

Avertissement

Les images qui illustrent les codes dans les tableaux ne sont pas normatives. Des variations considérables peuvent exister d'une police à l'autre.

Les tableaux qui suivent représentent les caractères de la norme ISO/CEI 10646 et du standard Unicode. Les caractères sont regroupés en blocs apparentés. Habituellement, ces blocs comprennent des caractères provenant d'une seule écriture. Un bloc de caractères suffit généralement à représenter une écriture. Il existe, toutefois, des exceptions notables, plus particulièrement dans le domaine des caractères de ponctuation.

Une liste de noms de caractère suit chaque tableau de codes, à l'exception des idéophonogrammes CJC et des syllabes hangûl ; pour plus de détails se référer aux sections *15.2 Idéophonogrammes unifiés CJC* et *15.3 Syllabes hangûl*. La liste de noms de caractère reprend tous les caractères du bloc et fournit le plus souvent des renseignements complémentaires sur ceux-ci.

On trouvera à la fin de ce livre un index reprenant les noms de caractères importants. La liste complète des noms de caractères se trouve dans la *Base de données des caractères Unicode* reprise sur le disque qui accompagne ce livre.

15.1 Liste des noms de caractère

L'exemple suivant illustre les différentes parties de chaque entrée de la liste de noms de caractère.

Code	*Image*	*Description*	
00AE	®	SYMBOLE MARQUE DÉPOSÉE	(nom officiel ISO 10646)
00AF	¯	MACRON	(nom officiel ISO 10646)
		= barre supérieure APL	(nom optionnel)
		•ce caractère chasse	(renseignement)
		→ 02c9 ˉ lettre modificative macron	(renvois)
		→ 0304 ◌̄ diacritique macron	
		→ 0305 ◌̅ diacritique tiret haut	
		≈ 0020 SP 0304 ◌̄	(décomposition de compatibilité)
00E5	å	LETTRE MINUSCULE LATINE A ROND EN CHEF	
		•danois, norvégien, suédois, wallon	
		≡ 0061 a 030A ◌̊	(décomposition canonique)

Images dans les tableaux de codes et dans les listes de caractères

Chaque caractère dans ces tableaux de codes est représenté à l'aide d'un glyphe représentatif. Ce glyphe n'a pas de valeur normative, il permet simplement à un utilisateur averti de reconnaître le caractère visé ou de retrouver facilement le caractère dans les tableaux. Dans de nombreux cas, il existe des représentations concurrentes, plus ou moins établies, pour un même caractère.

Les concepteurs de polices de haute qualité devront effectuer leur propre recherche pour déterminer l'apparence la plus appropriée des caractères Unicode. En outre, de nombreuses écritures nécessitent des formes de glyphes différentes selon le contexte, un positionnement contextuel du glyphe ou la formation de ligatures. Aucune de ces informations n'est illustrée dans les tableaux de codes.

Les glyphes représentatifs que l'on retrouve dans les tableaux sont des linéales, c'est-à-dire sans empattement, elles s'inspirent des polices Arial ou Helvetica. Ainsi, même le caractère ASCII U+0061 LETTRE MINUSCULE LATINE A se représente-t-il sous deux formes habituelles : « a » et « 𝒶 ». Dans un police à empattements comme Helvetica, le caractère U+03A5 LETTRE MAJUSCULE GRECQUE UPSILON ressemble à « Y », alors qu'on retrouve dans d'autres styles de police la forme « ϒ».

Un autre exemple : U+101F LETTRE MINUSCULE LATINE D CARON. Cette lettre se représente le plus souvent sous la forme « ď» plutôt que « ǆ». Dans ces cas-là, les tableaux représentent la variante la plus fréquente et non le modèle didactique.

On a unifié de nombreux caractères qui, dans de différentes langues, ont des apparences différentes. La forme qui illustre U+2116 № SYMBOLE NUMÉRO est un glyphe de pleine chasse comme on les rencontre dans les polices d'Extrême-Orient. En cyrillique, le glyphe reconnu par tous est « № ».

À maintes reprises, il a fallu représenter les caractères par des glyphes condensés, déplacés ou même déformés afin qu'ils respectent le format des tableaux de codes. Par exemple, U+0D10 ഐ LETTRE MALAYALAM AÏ est représentée sous une forme réduite afin de rentrer dans la cellule de tableau.

Il arrive parfois qu'il faille donner une forme artificielle aux caractères afin qu'on puisse les reconnaître dans les tableaux de codes. C'est le cas de U+00AD SHY TRAIT D'UNION VIRTUEL et de U+2011 NB TRAIT D'UNION INSÉCABLE où la fonction particulière du trait d'union est indiquée par la boîte en pointillé et les lettres.

Le contexte textuel des caractères fournit d'importantes indications sur leur identité, leur taille et le positionnement. Dans les tableaux de code, ces indices manquent. Ainsi, U+2075 ^{5} EXPOSANT CINQ est-il représenté en nettement plus petit que dans un texte composé avec un police similaire à du Times.

Les caractères combinatoires sont illustrés par rapport à un cercle en pointillé — par exemple, U+0940 ◌ी VOYELLE DIACRITIQUE DÉVANÂGARÎ ी. Le cercle en pointillé représente la position approximative du caractère de base. Lors du rendu, il est souvent nécessaire d'effectuer des ajustements supplémentaires. Les accents tels que 0302 ◌̂ DIACRITIQUE ACCENT CIRCONFLEXE doivent être ajustés verticalement et horizontalement selon la hauteur et la largeur du caractère de base, comparer « î » et « Ŵ ».

Pour les écritures non européennes, on a choisi comme styles de caractère ceux dont les œils se différenciaient le plus.

La norme ISO/CEI 10646 comprend de nombreux caractères utilisés par des langues mortes ou de moindre diffusion. Pour les langues où l'on ne dispose pas d'un corpus imprimé suffisant, il se peut que le dessin et la composition typographiques d'un caractère ne soient pas établis.

Renvois

Les caractères de renvoi (précédés de →) désignent plusieurs types de renvoi : l'inégalité explicite, les autres casses du même caractère ou d'autres rapports linguistiques.

Inégalité explicite. Les deux caractères ne sont pas identiques bien que leurs glyphes sont identiques ou fort similaires.

003A : DEUX-POINTS
→ 0589 : point arménien
→ 2236 : rapport

Autres rapports linguistiques. Parmi ces rapports, on compte les translittérations (entre le serbe et le croate, par exemple), des caractères sans rapport typographique mais qui servent à représenter le même son, etc.

01C9 lj LETTRE MINUSCULE LATINE LJ
= digramme soudé lj
→ 0459 љ lettre minuscule cyrillique lié
≈ 006C l 006A j

Transformations de casse

Quand on peut prévoir les différentes casses d'un caractère en se basant sur son nom et l'indication de casse (MINUSCULE ou MAJUSCULE), on ne retrouve pas le caractère correspondant à l'autre casse précisé en annotation. Cette correspondance est cependant explicitement précisée dans la *Base de données de caractères Unicode* présente sur le disque qui accompagne ce livre.

0041 A LETTRE MAJUSCULE LATINE A

01F2 Dz LETTRE MAJUSCULE LATINE D AVEC LETTRE MINUSCULE Z
≈ 0044 D 007A z

Lorsqu'il est impossible de prédire la correspondance de casse à partir du nom, on fournit alors cette information en annotation.

00DF ß LETTRE MINUSCULE LATINE S DUR
= eszett, s dur allemand
•allemand
•la majuscule est « SS »
•à l'origine une ligature de 017F ſ et 0073 s
→ 03B2 β lettre minuscule grecque bêta

Décompositions

La séquence de décomposition d'un caractère (constituée d'une ou plusieurs lettres) peut être de deux types : canonique ou de compatibilité. La correspondance canonique est indiquée à l'aide d'un symbole *identique à* ≡.

00E5 å LETTRE MINUSCULE LATINE A ROND EN CHEF
•danois, norvégien, suédois, wallon
≡ 0061 a 030A ◌̊

212B Å SYMBOLE ANGSTRÖM
≡ 00C5 Å lettre majuscule latine a rond en chef

La correspondance de compatibilité s'indique à l'aide d'un signe *presque égal à* ≈. Une balise de formatage peut accompagner la décomposition.

01F2 Dz LETTRE MAJUSCULE LATINE D AVEC LETTRE MINUSCULE Z
≈ 0044 D 007A z

FF21 A LETTRE MAJUSCULE LATINE A PLEINE CHASSE
≈ <large> 0041 A

On retrouve l'équivalent anglais des balises de formatage de compatibilité suivantes dans la *Base de données de caractères Unicode* :

<police>	Une variante de police (par exemple, une lettre de type gothique)
<insécable>	La version insécable d'une espace, d'un tiret ou d'une autre ponctuation
<initiale>	La forme de présentation initiale (arabe)
<médiale>	La forme de présentation médiale (arabe)
<finale>	La forme de présentation finale (arabe)
<isolée>	La forme de présentation isolée (arabe)
<cerclée>	Une forme cerclée
<exp>	Une forme suscrite ou en exposant
<souscrite>	Une forme souscrite
<verticale>	Une forme utilisée en composition verticale
<large>	Un caractère de compatibilité de pleine chasse (zenkaku)
<étroite>	Un caractère de compatibilité de demi-chasse (hankaku)
<petite>	Une petite variante de forme (compatibilité avec le jeu CNS 11643)
<enCarré>	Une variante de fonte CJC disposé en carré
<fraction>	Une fraction vulgaire
<compat>	Un caractère de compatibilité qui n'est pas autrement défini

La balise <compat> a été supprimée de la liste des noms de caractères qui accompagne les tableaux de codes, cependant toutes les autres balises de formatage de compatibilité se retrouvent explicitement mentionnées dans les décompositions de compatibilité.

Les décompositions mentionnées ne sont pas nécessairement complètes. Ainsi peut-on poursuivre la décomposition de U+0212B Å SYMBOLE ANGSTRÖM, en décomposant cette fois 00C5 Å LETTRE MAJUSCULE LATINE A ROND EN CHEF de manière canonique (pour plus de renseignements sur la décomposition, se reporter à la section *3.6, Décomposition*).

Renseignements sur les langues

Quand cela peut-être utile, on trouve parfois une note informative identifiant les langues qui utilisent ce caractère. Pour les lettres bicamérales, cette information n'est fournie que pour les lettres de bas de casse afin d'éviter une répétition inutile. Les points de suspension «…» indiquent que la liste des langues n'est pas limitative et qu'elle ne reprend que les langues principales.

Codes réservés

Les codes qualifiés d'un <*réservé*> n'ont été affectés à aucun caractère, ils sont réservés à une normalisation ultérieure. Les codes réservés se représentent à l'aide du glyphe ▨.

060D ▨ <réservé>

Les codes réservés peuvent renvoyer à des caractères qui sont codés ailleurs dans la norme.

2073 ▨ <réservé>
→ 00B3 ³ exposant trois

Les codes qualifiés d'un <*pas un caractère*> ne seront jamais affecté à un caractère. Ces codes se représentent à l'aide du glyphe ■.

FFFF ■ <pas un caractère>
on est assuré que la valeur FFFF ■ n'est en aucune façon un caractère Unicode

15.2 Idéogrammes unifiés CJC

Les blocs de caractères *Extension A aux idéogrammes unifiés CJC* et *Idéogrammes unifiés CJC* ne comportent pas de liste de noms de caractère puisque le nom d'un idéophonogramme unifié consiste simplement à ajouter sa valeur Unicode à la chaîne IDÉOGRAMME CJC UNIFIÉ- .

Comme pour les autres caractères des tableaux, chaque caractère de ces blocs est représenté à l'aide d'un glyphe représentatif et de sa valeur Unicode. Afin de se conformer aux différentes traditions typographiques qui existent en Extrême-Orient, il se peut qu'on doive utiliser des glyphes différents des ceux représentés dans les tableaux.

On retrouvera sur le disque qui accompagne cet ouvrage un tableau de correspondance entre les idéophonogrammes inclus dans la norme et ceux présents dans d'autres jeux de caractères normalisés.

15.3 Syllabes hangûl

La zone de syllabes hangûl (U+AC00..U+D7A3) ne comporte pas de liste de noms de caractères car le nom d'une syllabe coréenne peut être déterminé par l'application de l'algorithme décrit aux sections *3.11, Comportement des jamos jointifs* et *6.16, Hangûl*.