Préface

<< texte à écrire à la fin >>

0.1 Au sujet de la norme ISO/CEI 10646 et du standard Unicode

Ce livre définit la version 3.1 du standard Unicode. Les principes généraux, l'architecture du standard Unicode, les exigences de conformité et les conseils de mise en œuvre précèdent les informations liées au codage proprement dit. On trouvera dans les annexes d'utiles informations connexes. Le disque optique qui accompagne ce livre comprend des tableaux destinés aux développeurs ainsi que tous les rapports techniques publiés jusqu'à présent.

Concepts, architecture, conformité et lignes directrices

Les six premiers chapitres de la version 3.1 présentent le standard Unicode et fournissent les informations dont un développeur a besoin pour produire une mise en œuvre conforme. On y décrit les manipulations de texte de base, le traitement des signes combinatoires, les formes de codage et la disposition de textes bidirectionnels. Un chapitre particulier consacré aux conseils de mise en œuvre aborde de nombreuses questions qui se posent habituellement quand on implémente Unicode.

Le chapitre 1 présente les concepts de base du standard, les bases de l'architecture, le domaine d'application et les exigences en matière de manipulations de texte de base.

Le chapitre 2 formule les principes fondamentaux du standard Unicode, il illustre également des sujets précis comme les manipulations de texte, les propriétés générales de caractère et l'utilisation des caractères combinatoires.

Le chapitre 3 établit les exigences formelles de conformité. Ce chapitre présente également des algorithmes normatifs pour trois processus : l'ordonnancement canonique des signes combinatoires, le codage des syllabes coréennes hangûls à l'aide de jamos jointifs et, enfin, le formatage de textes bidirectionnels.

Le chapitre 4 décrit en détail les propriétés de caractères, à la fois normatives (obligatoires) et purement informatives. On retrouve sur le disque qui accompagne cet ouvrage des tableaux fournissant des renseignements supplémentaires sur les propriétés de caractères.

Le chapitre 5 aborde différentes questions relatives à la mise en œuvre de ce standard, parmi lesquelles la compression, les stratégies à adopter avec des caractères inconnus ou manquants ou le transcodage vers d'autres normes.

Le chapitre 6 spécifie quatre formes normalisées de texte Unicode. Grâce à celles-ci des textes équivalents possèdent des représentations binaires identiques.

Description des blocs de caractères

Le chapitre 7 décrit les caractères de ponctuation générale.

Le chapitre 8 présente les écritures alphabétiques européennes, y compris les écritures latine, grecque, cyrillique, arménienne, géorgienne, runique, ogamique et leurs signes combinatoires.

Le chapitre 9 présente les écritures de droite à gauche du Moyen-Orient : l'hébreu, l'arabe, le syriaque et le thâna.

Le chapitre 10 traite des écritures du Sud et du Sud-Est asiatique, parmi lesquelles le dévanâgarî, le bengali, le gourmoukhî, le goudjarati, l'oriya, le tamoul, le télougou, le kannara, le malayalam, le singhalais, le thaï, le lao, le tibétain et le birman.

Le chapitre 11 présente les écritures de l'Orient, y compris le han, l'hiragana, le katakana, le hangûl, le bopomofo et le yi.

Le chapitre 12 présente d'autres écritures : l'éthiopien, le chérokî, les syllabaires autochtones canadiens et le mongol.

Le chapitre 13 traite des symboles, parmi lesquels les devises, les caractères de type lettre, les symboles techniques et les opérateurs mathématiques.

Le chapitre 14 décrit les caractères spéciaux tels que ceux de la zone à usage privé, les seizets d'indirection et les caractères spéciaux proprement dits.

Tableaux

Le chapitre 15 reprend tous les caractères faisant partie du standard Unicode et fournit les renseignements suivants : leurs numéros de code, leurs noms, des renvois utiles vers d'autres caractères et des renseignements descriptifs importants.

Annexes et index

Les annexes contiennent des renseignements généraux sur des sujets importants : les systèmes de codage de caractères, l'historique d'Unicode et une comparaison avec son pendant ISO, l'ISO/CEI 10646.

L'annexe B reprend une bibliographie complète des normes et des ouvrages consultés pour rédiger cet ouvrage.

L'annexe C compare la norme internationale ISO/CEI 10646 et le standard Unicode.

L'annexe G reprend un glossaire explicatif des principaux termes utilisés dans le standard d'Unicode.

L'annexe H décrit l'histoire de l'unification han dans le standard Unicode.

Les annexes I comprennent deux index, l'index des noms Unicode qui permet de retrouver le numéro d'un caractère dont on connaîtrait le nom ou l'écriture et un index général qui renvoie aux sections correspondantes des textes normatifs et informatifs.

Enfin, *l'annexe L* reprend un lexique anglais – français des principaux termes du standard Unicode.

Base de données des caractères Unicode et rapports techniques

On désigne sous le nom de *Base de données des caractères Unicode* un ensemble de fichiers qui comprennent des valeurs de code de caractères, des noms de caractères et des données de propriété de caractère. Cette base de données est décrite en plus de détails dans le fichier UnicodeCharacterDatabase.html. La version 3.1 de la base de données est incluse sur le disque qui accompagne cet ouvrage. Des mises à jour et des révisions sont disponibles en ligne sur le site Unicode. Pour plus de renseignements sur les dernières versions disponibles, consultez :

http://www.unicode.org/unicode/standard/versions/

Les annexes normatives d'Unicode suivantes font officiellement partie de ce standard :

- UAX n° 9 : *The Bidirectional Algorithm*[1], version 3.1.0
- UAX n° 11 : *East Asian Width*, version 3.1.0
- UAX n° 13 : *Unicode Newline Guidelines*, version 3.1.0
- UAX n° 14 : *Line Breaking Properties*, version 3.1.0
- UAX n° 15 : *Unicode Normalization Forms*[2], version 3.1.0
- UAX n° 19 : *UTF-32*, version 3.1.0.

Les dernières versions de ces versions se retrouvent sur le cédérom. Des mises à jour et des révisions sont disponibles en ligne. Pour plus de renseignements au sujet des dernières versions, consultez http://www.unicode.org/unicode/standard/versions/.

Sur le disque

Le disque comprend la *Base de données des caractères Unicode*, elle fournit les numéros de caractère, les noms de caractère, les propriétés de caractère ainsi que les décompositions des caractères décomposables ou compatibles. En plus des rapports techniques Unicode et de la base de données de caractères Unicode qui font partie de ce standard, le disque comprend également d'autres rapports techniques (sur différents sujets comme la compression, le tri et les formats de transformation) ainsi que des tableaux de correspondance de propriétés (par exemple, des tableaux pour la casse) et des tableaux de conversion entre Unicode et des jeux de caractères internationaux, nationaux et de l'industrie (y compris des tableaux de renvois han). Veuillez vous référer au fichier LISEZ.MOI du disque pour une description complète du contenu.

0.2 Conventions de notation

Tout au long de ce livre, on utilise certaines conventions typographiques. Dans le corps du texte, une valeur de code Unicode se représente à l'aide d'un *U+n*, où *n* est un nombre composé de quatre à six chiffres en notation hexadécimale (en d'autres mots, il s'écrit à l'aide des chiffres 0 à 9 et des lettres A à F pour représenter les chiffres 10 à 15). Les zéros initiaux

[1] Le texte normatif de cette annexe correspond, à quelques paragraphes introductifs près, à la section 3.12 de ce livre.
[2] Cette annexe normative constitue le chapitre 6 de cet ouvrage.

doivent être supprimés pour autant que le nombre comporte toujours au moins quatre chiffres. Example : U+0001, U+0012, U+0123, U+1234, U+12345, U+102345.

- U+0416 est la valeur Unicode qui représente le caractère Ж dénommé LETTRE MAJUSCULE CYRILLIQUE JÉ.

À des fins de concision, on omet parfois le *U+* dans les tableaux.

Un intervalle de valeurs Unicode s'exprime de la façon suivante : *U+xxxx→U+yyyy ou U+xxxx—U+yyyy* ou encore *U+xxxx…U+yyyy* ; *xxxx* et *yyyy* représentent respectivement la première et la dernière valeur Unicode de l'intervalle. La flèche, le tiret sur cadratin et les points indiquent qu'il s'agit d'un intervalle continu qui inclut les deux extrémités.

- L'intervalle U+0900→U+097F comprend 128 valeurs de caractère.

Tous les caractères Unicode de ce livre ont un nom unique qui correspond à ceux de la version française de la norme internationale ISO/CEI 10646. Les noms de caractère Unicode se composent de lettres majuscules latines A à Z, de lettres accentuées de l'ISO/CEI 8859-15, de chiffres, de l'espace, de l'apostrophe ou du trait-d'union-signe-moins. Cette convention facilite la génération automatique d'identificateurs pour langage informatique à partir des noms. Les idéogrammes unifiés d'Orient portent le nom de `IDÉOGRAMME CJC UNIFIÉ-X`, où X correspond à leur valeur Unicode hexadécimale, par exemple `IDÉOGRAMME CJC UNIFIÉ-4E00`. Les noms des syllabes hangûl se déduisent algorithmiquement ; pour plus détails voir Noms de syllabes hangûl dans la section *3.11, Comportement des jamos jointifs.*

Dans le corps du texte, le nom officiel des caractères Unicode apparaît en petites capitales (par exemple, LETTRE MINUSCULE GRECQUE MU), les noms facultatifs (les alias) apparaissent en italique (par exemple, *ligature ij*). On utilise également l'italique pour désigner un élément textuel qui n'est pas codé explicitement (par exemple, *alef pasekh*) ou pour faire ressortir un mot étranger (par exemple, le mot gallois *ynghyd*). Les transcriptions phonologiques se font entre barres obliques (ou *cotices*), comme dans la transcription du khmer /khnyom/.

Le chapitre *15, Tableaux de codes*, commence par une description des symboles utilisés dans la liste des noms de caractères.

Dans le texte de ce livre, le nom Unicode utilisé seul désigne le standard Unicode.

BNF étendue

Le standard Unicode et les rapports techniques Unicode utilisent comme notation syntaxique une forme de BNF étendue. Étant donné qu'il existe différentes conventions de BNF, le tableau *0-1, BNF étendue,* précise la notation utilisée dans cet ouvrage.

Dans le corps du texte, on représente une suite de *numéros de caractères* (ou *points de code,* voir le *Glossaire*) par une liste entre crochets angulaires dont les éléments sont séparés à l'aide d'une virgule. On utilise ici U+003C < SIGNE INFÉRIEUR À et U+003E > SIGNE SUPÉRIEUR À comme crochets angulaires. La virgule peut être facultativement suivie d'espaces. On nomme une suite notée de la sorte, un *identificateur de suite Unicode* (ISU).

Quand un contexte clair le permet, on peut également représenter une suite de caractères à l'aide de noms abrégés : <*a circonflexe*>. Il est alors permis d'omettre les crochets angulaires.

Contrairement aux suites de *points de code*, on peut représenter une suite d'un ou plusieurs *unités de stockage* à l'aide d'une liste entre crochets angulaires, mais dont les éléments ne

sont pas séparés par une virgule et n'utilisent pas la notation U+ . Ainsi, la notation <*nn nn nn nn*> représente une suite d'octets, tels que ceux utilisés pour décrire un caractère Unicode dans la forme de stockage UTF-8. La notation <*nnnn nnnn*> représente, pour sa part, une suite de seizets, comme ceux utilisés pour stocker un caractère Unicode sous la forme de stockage UTF-16.

Dans d'autres environnements, comme ceux des langages de programmation ou de balisage, on peut utiliser une autre notation pour représenter une suite de numéros de caractère ou d'unités de stockage.

Classes de caractères. Une classe de caractères se compose à l'aide d'un ou deux ensembles de caractères de base. Elle correspond soit à un seul ensemble de base, à la négation d'un ensemble de base ou à la différence (ensembliste) entre deux ensembles de base. Les ensembles de base sont délimités par des crochets, ils énumèrent des listes des caractères, des intervalles de caractères, des catégories générales ou des négations de catégorie générale. On trouvera la syntaxe correspondante ci-dessous :

Tableau 0-1. BNF étendue

Symboles	Signification
`x := …`	règle de production
`x y`	suite constituée de x puis d'y
`x*`	x apparaît 0 ou plusieurs fois
`x?`	x apparaît 0 ou 1 fois
`X+`	x apparaît au moins 1 fois
`x | y`	x ou y
`( x )`	groupement
`x || y`	équivalent de `(x | y | (x y ))`
`{ x }`	équivalent de `(x)?`
`"abc"`	littéral chaîne de caractères (on désigne parfois l'espace par un `"_"` pour plus de clarté)
`'abc'`	littéral chaîne de caractères (forme concurrente)
`\u1234`	caractère Unicode au sein d'une chaîne ou d'une classe de caractères
`\v00101234`	valeur scalaire Unicode au sein d'une chaîne ou d'une classe de caractères
`U+HHHH`	littéral caractère Unicode : équivalent à '\uHHHH'
`U-HHHHHHHH`	littéral caractère Unicode : équivalent à '\vHHHHHHHH'
`classeCar`	classe de caractère (syntaxe ci-dessous)

```
classeCar  := ensDeBase | '¬' ensDeBase | ensDeBase '-' ensDeBase
ensDeBase  := '[' élément ( ','? élément)* ']'
élément    := car | car '-' car |  '{' '¬'? catégorie '}'
```

Les catégories générales se trouvent au chapitre *4, Propriétés des caractères*. Exemple : `[{Lettre Majuscule}]` pour désigner les *majuscules*. Les catégories principales comme `[{Signes}]` sont équivalentes à une liste de sous-catégories : `[{Signe à chasse`

`nulle}{Signe avec chasse}{Signe englobant}]`. On trouve quelques exemples supplémentaires ci-dessous dans le tableau 0-2, Exemples de classe de caractères.

Tableau 0-2. Exemples de classe de caractères

Syntaxe	Correspond à
`[a-z]`	lettres minuscules latines non accentuées
`[a-z,â,à,ä,ç,é,ê,è,ë,ï,î,ù,û,ü,ô,ö,ÿ]`	lettres minuscules françaises
`¬[c]`	toutes les lettres sauf *c*
`[0-9]`	chiffres décimaux européens
`[\u0030-\u0039]`	(*idem* mais en utilisant des échappements Unicode)
`[0-9,A-F,a-f]`	chiffres hexadécimaux
`[{Lettre},{Signe sans chasse}]`	toutes les lettres et signes sans chasse
`[{L},{Mn}]`	(idem mais avec une notation abrégée)
`[{¬Cn}]`	tous les caractères Unicode affectés
`[\u0600-\u06FF]-[{Cn}]`	tous les caractères arxabes affectés

Opérateurs

Le tableau 0-3 reprend la liste des opérateurs utilisés dans ce standard.

Tableau 0-3. Opérateurs

÷	coupure permise ici (voir section *5.15, Repérage des frontières d'élément textuel)*
×	position insécable
→	est transformé en ou se comporte comme
⁄	division entière (arrondie)
%	modulo (équivalent au reste entier de la division pour les nombres positifs)

Texte grisé

Les cinq premiers chapitres de cet ouvrage sont une traduction de la version anglaise du standard Unicode, version 3.1. Le chapitre six correspond à l'annexe normative d'Unicode n°15. Le texte anglais de ces six chapitres est traduit le plus fidèlement possible, nos annotations sont en note de bas de page ou en grisé dans le corps du texte lorsqu'une annotation aurait été trop longue. Les autres chapitres et annexes sont pour la plupart des adaptations du texte du standard Unicode, certains sont totalement originaux, leur texte n'apparaît pas en grisé.

0.3 Ressources

Il existe un certain nombre de ressources en ligne où l'on trouvera un complément de renseignements et de données sur le standard Unicode, ainsi que des mises à jours et des corrections. En voici une brève liste :

Site internet en français

- http://hapax.iquebec.com

Site internet Unicode en anglais

- http://www.unicode.org

Forum « Usenet » en français

Il existe également un forum « Usenet » francophone se rapportant à Unicode, il s'agit de `fr.comp.normes.unicode` .

Comment contacter le JTC1/SC2/GT2

Pour toute commande de normes ou renseignement ou suggestion concernant les normes de jeux de caractères, les lecteurs sont invités à contacter leur représentant national en premier lieu :

- Belgique (IBN)
  Institut belge de normalisation
  Av. de la Brabançonne 29
  1000 BRUXELLES

  Téléphone : +32 2 738 01 11
  Télécopie : +32 2 733 42 64
  Courriel : croon@ibn.be

- Canada (SCC)
  Conseil canadien des normes
  270 rue Albert, Bureau 200
  Ottawa (Ontario)
  K1P 6N7

  Téléphone : +1 613 238 32 22
  Télécopie : +1 613 995 45 64
  Courriel : isosd@scc.ca
  Internet : http://www.scc.ca/homef.html

- France (AFNOR)
  Association française de normalisation
  Tour Europe
  92049
  PARIS
  LA DÉFENSE CEDEX
  Téléphone : +33 1 42 91 55 55
  Télécopie : +33 1 42 91 56 56
  Courriel : international@email.afnor.fr
  Internet : http://www.afnor.fr/

- Maroc (SNIMA)

Service de normalisation industrielle marocaine
Ministère du commerce, de l'industrie et l'artisanat
Angle Avenue Al Filao et rue Dadi
Secteur 21 Hay Riad
10100 RABAT
Téléphone : +212 7 71 62 15
Télécopie : +212 7 71 17 98
Courriel : snima@mcinet.gov.ma
Internet : http://www.mcinet.gov.ma

- Tunisie (INNORPI)
  Institut national de la normalisation et de la propriété industrielle
  B.P. 23
  1012 TUNIS-BELVÉDÈRE
  Téléphone : +216 1 78 59 22
  Télécopie : +216 1 78 15 63
  Courriel : inorpi@email.ati.tn

Les lecteurs dont le comité national n'est pas mentionné ci-dessus peuvent toutefois rentrer en contact avec leur organisme national chargé de la normalisation. Si celui-ci ne participe pas aux travaux du JTC1/SC2/GT2 ou ne disposent pas des normes recherchées, ils pourront encore rentrer directement en contact avec l'Organisation internationale de normalisation (ISO) à Genève (pour la commande de normes, par exemple) ou le secrétariat du JTC1/SC2/GT2 pour des questions portant sur les normes de jeux de caractères.

- Organisation internationale de normalisation (ISO)
  1, rue de Varembé
  Case postale 56
  CH-1211 Genève 20
  Suisse

  Téléphone : +41 22 749 01 11
  Télécopie : +41 22 733 34 30
  Courriel : central@iso.ch
  Internet : http://www.iso.ch/indexf.html

- Secrétaire du JTC1/SC2/GT2
  a/s IPSJ/ITSCJ
  Kikai Shinko Building
  3-5-8 Shibakoen, Minato-Ku
  Tôkyô 105
  Japon

  Téléphone : +81 3 34 31 28 08
  Télécopie : +81 3 34 31 64 93
  Courriel : kimura@itscj.ipsj.or.jp