NORME INTERNATIONALE

ISO/CEI 10646-1

Deuxième édition
2000-09-15

Technologies de l'information — Jeu universel de caractères codés sur plusieurs octets (JUC) —

Partie 1: Architecture et plan multilingue de base

Information technology — Universal Multiple-Octet Coded Character Set (UCS) —

Part 1: Architecture and Basic Multilingual Plane

Numéro de référence
ISO/CEI 10646-1:2000(F)

ISO copyright office
Case postale 56 • CH-1211 Geneva 20
Tel. + 41 22 749 01 11
Fax. + 41 22 749 09 47
E-mail copyright@iso.ch
Web www.iso.ch

Imprimé en Suisse

Sommaire

Avant-propos

L'ISO (Organisation internationale de normalisation) et la CEI (Commission électrotechnique internationale) forment le système spécialisé de la normalisation mondiale. Les organismes nationaux membres de l'ISO ou de la CEI participent au développement de Normes internationales par l'intermédiaire des comités techniques créés par l'organisation concernée afin de s'occuper des domaines particuliers de l'activité technique. Les comités techniques de l'ISO et de la CEI collaborent dans des domaines d'intérêt commun. D'autres organisations internationales, gouvernementales et non gouvernementales, en liaison avec l'ISO et la CEI participent également aux travaux.

Les Normes internationales sont rédigées conformément aux règles données dans les Directives ISO/CEI, Partie 3.

Dans le domaine des technologies de l'information, l'ISO et la CEI ont créé un comité technique mixte, l'ISO/CEI JTC 1. Les projets de Normes internationales adoptés par le comité technique mixte sont soumis aux organismes nationaux pour vote. Leur publication comme Normes internationales requiert l'approbation de 75 % au moins des organismes nationaux votants.

L'attention est appelée sur le fait que certains des éléments de la présente partie de l'ISO/CEI 10646 peuvent faire l'objet de droits de propriété intellectuelle ou de droits analogues. L'ISO et la CEI ne sauraient être tenues pour responsables de ne pas avoir identifié de tels droits de propriété et averti de leur existence.

La Norme internationale ISO/CEI 10646-1 a été élaborée par le comité technique mixte ISO/CEI JTC 1, *Technologies de l'information*, sous-comité SC 2, *Jeux de caractères codés*.

Cette deuxième édition annule et remplace la première édition (ISO/CEI 10646-1:1993), qui a fait l'objet d'une révision technique. Elle incorpore aussi les Amendements 1 à 13, 16 à 21, et 23, ainsi que les Rectificatifs techniques 1 et 2, relatifs à la première édition.

L'ISO/CEI 10646 comprend les parties suivantes, présentées sous le titre général *Technologies de l'information — Jeu universel de caractères codés sur plusieurs octets (JUC)*:

— *Partie 1: Architecture et plan multilingue de base*

— *Partie 2: Plan multilingue secondaire pour caractères et symboles, plan supplémentaire pour idéogrammes CJK, plan à but particulier*

Des parties complémentaires définiront d'autres plans.

Les annexes A à D constituent des éléments normatifs de la présente partie de l'ISO/CEI 10646. Les annexes E à S sont données uniquement à titre d'information.

Introduction

L'ISO/CEI 10646 normalise le jeu universel de caractères codés sur plusieurs octets (JUC). Elle s'applique à la représentation, à la transmission, à l'échange, au traitement, à la sauvegarde, à la saisie et à la présentation des langues du monde sous forme écrite et de symboles complémentaires.

La présente partie de l'ISO/CEI 10646 traite de l'architecture globale et du plan multilingue de base du JUC.

Technologies de l'information — Jeu universel de caractères codés sur plusieurs octets (JUC) —

Partie 1:
Architecture et plan multilingue de base

1 Domaine d'application

L'ISO/CEI 10646 normalise le jeu universel de caractères codés sur plusieurs octets (JUC). Elle s'applique à la représentation, à la transmission, à l'échange, au traitement, au stockage, à la saisie et à la présentation des langues du monde sous forme écrite et de symboles complémentaires.

La présente partie de l'ISO/CEI 10646 traite de l'architecture générale et

- définit les termes utilisés dans l'ISO/CEI 10646 ;
- décrit la structure générale du jeu de caractères codés ;
- décrit le plan multilingue de base (PMB) du JUC et définit un ensemble de caractères graphiques utilisés dans la forme écrite des langues à l'échelle mondiale ;
- nomme et établit la représentation codée des caractères graphiques du PMB ;
- prescrit la forme canonique à quatre octets (32 bits) du JUC : UCS-4 ;
- précise une forme du PMB à deux octets (16 bits) pour le JUC : UCS-2 ;
- établit la représentation codée des fonctions de commandes ;
- établit la gestion de tout développement ultérieur du présent jeu de caractères codés.

Le JUC est un système de codage différent de celui décrit dans l'ISO/CEI 2022. La méthode employée pour désigner le JUC à partir de l'ISO/CEI 2022 est précisée en 16.2.

NOTE 1 : Le standard Unicode, version 3.0, établit un jeu de caractères et des représentations codées identiques à ceux de la partie 1 de la présente norme internationale. Il fournit également des précisions sur les propriétés des caractères ainsi que des algorithmes de traitement et des définitions utiles aux développeurs de logiciels.

NOTE 2 : Il est prévu que des positions de code de caractère pour des écritures supplémentaires soient attribuées dans la partie 1 de la présente norme internationale après réception de suffisamment de données et analyse par les organismes de normalisation nationaux ou des experts compétents.

2 Conformité

2.1 Généralités

En cas d'utilisation de caractères à usage privé, telle que précisée dans l'ISO/CEI 10646, les présentes exigences de conformité ne s'appliquent pas à ces caractères.

2.2 Conformité de l'échange d'information

Une donnée sous forme de caractères codés (donnée CC), au sein d'une information codée destinée à être échangée, est en conformité avec l'ISO/CEI 10646 si

a) toutes les représentations codées des caractères graphiques de la donnée CC sont conformes aux articles 6 et 7, à une forme identifiée choisie dans l'article 13, dans l'annexe C ou dans l'annexe D et à un niveau de mise en œuvre identifié choisi selon l'article 14 ;

b) tous les caractères graphiques représentés dans cette donnée CC proviennent d'un sous-ensemble identifié (article 12) ;

c) toutes les représentations codées des fonctions de commande dans cette donnée CC sont conformes à l'article 15.

Une déclaration de conformité doit identifier la forme adoptée, le niveau de mise en œuvre adopté et, au moyen d'une liste de collections ou de caractères, le sous-ensemble adopté.

2.3 Conformité des dispositifs

Un dispositif est en conformité avec l'ISO/CEI 10646 s'il satisfait aux exigences du point a) ci-dessous, ainsi qu'à celles du point b) ou du point c) ou des deux.

NOTE : Le terme dispositif est défini (en 4.12) comme élément d'un matériel de traitement de l'information qui peut transmettre ou recevoir des informations codées dans des données CC. Un dispositif peut être une unité d'entrée-sortie

classique ou un processus tel qu'un programme d'application ou une fonction passerelle.

Une déclaration de conformité doit identifier le document contenant la description mentionnée en a) ci-dessous ainsi que la ou les formes adoptées, le niveau de mise en œuvre adopté et, au moyen d'une liste de collections ou de caractères, le sous-ensemble et la liste des fonctions de commande adoptés selon l'article 15.

a) Description d'un dispositif : un dispositif conforme à l'ISO/CEI 10646 doit faire l'objet d'une description identifiant les moyens permettant à l'utilisateur de fournir des caractères au dispositif ou de les reconnaître lorsqu'ils sont mis à sa disposition, comme précisé respectivement aux paragraphes b) et c) ci-dessous.

b) Dispositif d'émission : un dispositif d'émission doit permettre à son utilisateur de fournir tous les caractères d'un sous-ensemble adopté et être capable de transmettre leurs représentations codées dans une donnée CC conformément à la forme et au niveau de mise en œuvre adoptés.

c) Dispositif de réception : un dispositif de réception doit être capable de recevoir et d'interpréter toute représentation codée des caractères d'une donnée CC conformément à la forme et au niveau de mise en œuvre adoptés et doit mettre à la disposition de l'utilisateur, de manière à lui permettre de les identifier, tous les caractères correspondants du sous-ensemble adopté.

Tous les caractères qui ne sont pas dans le sous-ensemble adopté doivent être signalés à l'utilisateur. La façon de les lui signaler ne doit pas nécessairement permettre de les distinguer les uns des autres.

NOTE 1 : L'indication fournie à l'utilisateur peut consister à rendre disponible un caractère particulier pour représenter tous ceux qui ne sont pas dans le sous-ensemble adopté ou à fournir un signal audible ou visible distinct adapté au type d'utilisateur.

NOTE 2 : Voir également l'annexe J pour les dispositifs de réception ayant une possibilité de retransmission.

3 Références normatives

Les documents normatifs suivants contiennent des dispositions qui, par suite de la référence qui y est faite, constituent des dispositions valables pour la présente partie de l'ISO/CEI 10646. Pour les références datées, les amendements ultérieurs ou les révisions de ces publications ne s'appliquent pas. Toutefois, les parties prenantes aux accords fondés sur la présente partie de l'ISO/CEI 10646 sont invitées à rechercher la possibilité d'appliquer les éditions les plus récentes des documents normatifs indiqués ci-après. Pour les références non datées, la dernière édition du document normatif en référence s'applique. Les membres de l'ISO et de la CEI possèdent le registre des Normes internationales en vigueur.

ISO/CEI 2022:1994, *Technologies de l'information — Structure de code de caractères et techniques d'extension.*

ISO/CEI 6429:1992, *Technologies de l'information — Fonctions de commande pour les jeux de caractères codés.*

4 Termes et définitions

Pour les besoins de la présente partie de l'ISO/CEI 10646, les termes et définitions suivants s'appliquent:

4.1 bloc : collection contiguë de caractères partageant des caractéristiques communes, par exemple l'appartenance à un système d'écriture. Un bloc n'en chevauche pas un autre. Une ou plusieurs positions de code au sein d'un bloc peuvent n'être associées à aucun caractère.

4.2 caractère : élément d'un ensemble utilisé pour organiser, commander ou représenter des données.

4.3 caractère codé : un caractère et sa représentation codée.

4.4 caractère combinatoire : élément d'un sous-ensemble identifié du jeu de caractères codés de l'ISO/CEI 10646 destiné à se combiner avec le caractère graphique non-combinatoire précédent, ou avec une suite de caractères combinatoires précédée d'un caractère non-combinatoire (voir également 4.34).

NOTE : La présente partie de l'ISO/CEI 10646 définit plusieurs groupes de sous-ensembles comprenant des caractères combinatoires.

4.5 caractère de compatibilité : caractère graphique inclus comme caractère codé de l'ISO/CEI 10646 principalement pour assurer la compatibilité avec des jeux de caractères codés existants.

4.6 caractère graphique : caractère, autre qu'une fonction de commande, qui a une représentation visuelle normalement manuscrite, imprimée ou affichée.

4.7 cellule : place dans une rangée à laquelle un caractère isolé peut être affecté.

4.8 collection : un ensemble de caractères codés qui est numéroté et nommé et qui comprend les caractères codés dont les positions de codes sont comprises dans les intervalles spécifiés.

NOTE : Si un des intervalles spécifiés comprend des positions de code auxquelles aucun caractère n'a été associé, le répertoire de cette collection changera si un nouveau caractère venait à être affecté à une de ces positions par une modification de cette norme internationale. Cependant, il est prévu que le numéro de la collection et son nom ne changeront pas dans les prochaines éditions de la présente norme internationale.

4.9 collection fixe : une collection où chaque position de code dans le ou les intervalles spécifiés est associée à un caractère et qui devrait rester inchangée dans les prochaines éditions de cette norme internationale.

4.10 demi-zone basse : un ensemble de cellules réservées pour utilisation par UTF-16 (voir annexe C) ; un seizet correspondant à l'une de ces cellules peut être le second d'une paire de seizets représentant un caractère d'un plan autre que le PMB.

4.11 demi-zone haute : un ensemble de cellules réservées pour utilisation par UTF-16 (voir annexe C) ; un seizet correspondant à l'une de ces cellules peut être le premier d'une paire de seizets représentant un caractère d'un plan autre que le PMB.

4.12 dispositif : élément d'un matériel de traitement de l'information qui peut transmettre ou recevoir des informations codées dans des données CC (il peut s'agir d'une unité d'entrée-sortie au sens classique ou d'un processus tel qu'un programme d'application ou une fonction passerelle).

4.13 donnée CC (donnée sous forme de caractères codés) : élément d'une information échangée, composé d'une suite de représentations codées de caractères, conformément à une ou plusieurs normes identifiées de jeux de caractères codés.

4.14 échange : transfert de données de caractères codés d'un utilisateur à un autre, en utilisant des moyens de télécommunication ou des supports interchangeables.

4.15 écriture : ensemble de caractères graphiques utilisé dans la forme écrite d'une ou de plusieurs langues.

4.16 état implicite : état présumé lorsque aucun état n'a été explicitement retenu.

4.17 fonction de commande : opération qui affecte l'enregistrement, le traitement, la transmission ou l'interprétation des données et qui a une représentation codée formée d'un ou de plusieurs octets.

4.18 forme canonique : forme de représentation d'un caractère du JUC utilisant quatre octets.

4.19 forme de présentation : dans la présentation de certaines écritures, forme du symbole graphique représentant un caractère en fonction de la position de ce caractère par rapport aux autres.

4.20 frontière de caractères : limite, dans une chaîne d'octets, entre le dernier octet de la représentation codée d'un caractère et le premier octet de celle du caractère codé suivant.

4.21 groupe : subdivision de l'espace de codage du présent jeu de caractères codés, formé de 256 × 256 × 256 cellules.

4.22 interopérabilité : processus permettant à deux ou plusieurs systèmes utilisant chacun des jeux de caractères codés différents d'échanger des données constituées de caractères codés ; il peut y avoir conversion entre les codes deux à deux.

4.23 jeu de caractères codés : ensemble de règles univoques qui définissent un groupe de caractères et établissent une correspondance entre chaque caractère et sa représentation codée.

4.24 octet : suite ordonnée de huit bits considérée comme une unité.

4.25 plan : subdivision d'un groupe ; se composant de 256 x 256 cellules.

4.26 plan à usage privé : plan du présent jeu de caractères codés dont le contenu n'est pas prescrit par l'ISO/CEI 10646 (voir article 10).

4.27 plan multilingue de base (PMB) : plan 00 du groupe 00.

4.28 plan supplémentaire : plan admettant des caractères qui n'ont pas été affectés au plan multilingue de base.

4.29 présentation ; présenter : opération d'écriture, d'impression ou d'affichage d'un symbole graphique.

4.30 rangée : subdivision d'un plan composée de 256 cellules.

4.31 répertoire : ensemble précis de caractères représentés dans un jeu de caractères codés.

4.32 seizet (élément RC) : une suite de deux octets comprenant l'octet R et l'octet C (voir 6.2) de la suite de 4 octets (dans la forme canonique) correspondant à une cellule de l'espace de codage de ce jeu de caractères codés.

4.33 seizet non apparié : seizet dans une donnée CC qui est soit :

- un seizet de la demi-zone haute qui n'est pas immédiatement suivi d'un seizet de la demi-zone basse ;
- un seizet de la demi-zone basse qui n'est pas immédiatement précédé d'un seizet de la demi-zone haute.

4.34 séquence composite : suite de caractères graphiques se composant d'un caractère non-combinatoire suivi d'un ou de plusieurs caractères combinatoires (voir également 4.4).

NOTE 1 : Le symbole graphique d'une séquence composite est généralement composé de la combinaison des symboles graphiques de chaque caractère dans la séquence.

NOTE 2 : Une séquence composite n'est pas un caractère et ne fait donc pas partie du répertoire de l'ISO/CEI 10646.

4.35 symbole graphique : représentation visuelle d'un caractère graphique ou d'une séquence composite.

4.36 tableau de code : tableau indiquant les caractères affectés aux octets d'un code.

4.37 tableau de code détaillé : tableau de code indiquant les caractères isolés et couvrant normalement une partie de rangée.

4.38 utilisateur : personne ou autre entité recourant au service assuré par le dispositif. (Cette entité peut être un processus tel qu'un programme d'application si « le dispositif » est un convertisseur de code ou une fonction passerelle, par exemple).

4.39 zone : suite de cellules d'un tableau de code comportant une ou plusieurs rangées, complètes ou partielles, contenant des caractères d'une catégorie particulière (voir article 8).

5 Structure générale du JUC

La structure générale du jeu universel de caractères codés sur plusieurs octets (dénommé ci-après « le présent jeu de caractères codés ») est décrite dans le présent article explicatif et illustrée par les figures 1 et 2. La description normative de la structure est donnée dans les articles suivants.

La valeur de chaque octet est exprimée en notation hexadécimale de 00 à FF dans l'ISO/CEI 10646 (voir l'annexe K).

La forme canonique du présent jeu de caractères codés – la manière dont il doit être conçu – utilise un espace de codage à quatre dimensions considéré comme une entité unique composée de 128 groupes à trois dimensions.

NOTE : Ainsi, le bit 8 de l'octet de poids fort dans la forme canonique d'un caractère codé peut être utilisé pour les besoins du traitement interne dans un dispositif tant qu'il a la valeur zéro dans une donnée CC conforme.

Chaque groupe est composé de 256 plans à deux dimensions et chaque plan de 256 rangées à une dimension, chaque rangée contenant 256 cellules. Un caractère est situé et codé au niveau d'une cellule dans cet espace de codage ; une cellule peut être déclarée inutilisée.

Sous la forme canonique, les quatre octets utilisés pour représenter chaque caractère correspondent respectivement au groupe, au plan, à la rangée et à la cellule. La forme canonique est composé de quatre octets car, d'une part, deux octets sont insuffisants pour couvrir tous les caractères possibles et, d'autre part, une représentation sur 32 bits s'accorde avec les architectures des processeurs modernes.

La forme canonique à quatre octets peut être utilisée comme jeu de caractères codés à quatre octets et s'appelle alors UCS-4.

Le premier plan (plan 00 du groupe 00) est appelé plan multilingue de base. Ce plan comporte des caractères usuels dans les écritures alphabétiques, syllabiques et idéographiques ainsi que divers chiffres et symboles.

Les plans ci-dessous sont considérés comme des plans supplémentaires ou à usage privé admettant d'autres caractères graphiques (voir l'article 9).

Les plans réservés à l'usage privé sont précisés à l'article 10. Le contenu des cellules des zones à usage privé n'est pas précisé dans l'ISO/CEI 10646.

L'emplacement de chaque caractère dans le JUC est fonction de ses octets de groupe, de plan, de rangée et de cellule.

En plus de la forme canonique, on définit une forme du PMB à deux octets. Le plan multilingue de base peut ainsi être utilisé comme jeu de caractères codés à deux octets dont l'identification est UCS-2.

Des sous-ensembles de l'espace de codage peuvent être utilisés afin d'obtenir un sous-répertoire de caractères graphiques.

Un format transformé du JUC (UTF-16) est décrit à l'annexe C ; il peut être utilisé pour représenter des caractères de 16 plans du groupe 00, en sus du PMB, sous une forme compatible avec la forme à deux octets du PMB.

Un format transformé du JUC (UTF-8) est décrit à l'annexe D ; il peut être utilisé pour transmettre des données textuelles par des systèmes de communication utilisant les valeurs d'octets correspondant aux caractères de commande codés selon la structure à 8 bits de l'ISO/CEI 2022 et selon

l'ISO/CEI 4873. UTF-8 évite aussi l'utilisation de valeurs d'octet selon l'ISO/CEI 4873 qui ont une signification particulière lors du traitement de noms de fichiers dans des systèmes de fichiers courants.

6 Structure de base et nomenclature

6.1 Structure

Le jeu universel de caractères codés sur plusieurs octets défini par l'ISO/CEI 10646 doit être perçu comme une seule entité.

L'ensemble du présent jeu de caractères codés doit être considéré comme formé de 128 groupes de 256 plans. Chaque plan est formé de 256 rangées de caractères, chaque rangée contenant 256 cellules. Dans un tableau de code représentant le contenu d'un plan (comme à la figure 2), l'axe horizontal doit représenter l'octet de poids faible, avec sa plus petite valeur à gauche, et l'axe vertical doit représenter l'octet de poids fort avec sa plus petite valeur en haut.

Un octet doit coder chaque axe de l'espace de codage. Dans chaque octet, le bit de poids fort sera le bit 8 et le bit de poids faible le bit 1.

En conséquence, le poids attribué à chaque bit doit être :

bit 8	bit 7	bit 6	bit 5	bit 4	bit 3	bit 2	bit 1
128	64	32	16	8	4	2	1

6.2 Codage des caractères

Dans la forme canonique du jeu de caractères codés, chaque caractère du jeu complet de caractères codés doit être représenté par une suite de quatre octets. L'octet de poids fort de cette suite doit être l'octet de groupe, l'octet de poids faible devant être l'octet de cellule. Cette suite peut donc être représentée ainsi :

poids fort poids faible

octet de groupe	octet de plan	octet de rangée	octet de cellule

Pour plus de concision, les octets peuvent être désignés comme suit :

poids fort poids faible

octet G	octet P	Octet R	octet C

Le cas échéant, on peut encore les abréger en G, P, R et C.

La valeur de chaque octet doit être représentée par deux chiffres hexadécimaux, par exemple : 31 ou FE. Si un seul caractère doit être identifié en termes des valeurs de ses groupe, plan, rangée et cellule, la représentation doit être la suivante :

0000 0030 pour CHIFFRE ZÉRO

0000 0041 pour LETTRE MAJUSCULE LATINE A

Pour désigner des caractères dans un plan, on peut supprimer les quatre premiers zéros (pour les octets G et P). Par exemple, 0030 peut être utilisé pour désigner CHIFFRE ZÉRO.

6.3 Ordre des octets

La suite d'octets représentant un caractère ainsi que ses terminaisons de poids fort et de poids faible doivent suivre l'ordre et la position indiqués ci-dessus. Dans le cas d'une mise en série sous forme d'octets, un octet de poids fort doit précéder les octets de poids faible. En l'absence de mise en série sous forme d'octets, l'ordre des octets peut être convenu entre l'émetteur et le destinataire (voir 16.1 et l'annexe H).

6.4 Choix des noms des caractères

Chaque caractère graphique de l'ISO/CEI 10646 est identifié par un nom unique pour une version en une langue donnée. Le nom d'un caractère doit :

a. ou refléter sa signification habituelle ;
b. ou décrire la forme du symbole graphique correspondant ;
c. ou suivre la règle de l'article 27 dans le cas des idéogrammes unifiés chinois/japonais/coréens.

Des conseils pour la création des noms de caractères dans les cas a. et b. ci-dessus sont donnés à l'annexe L.

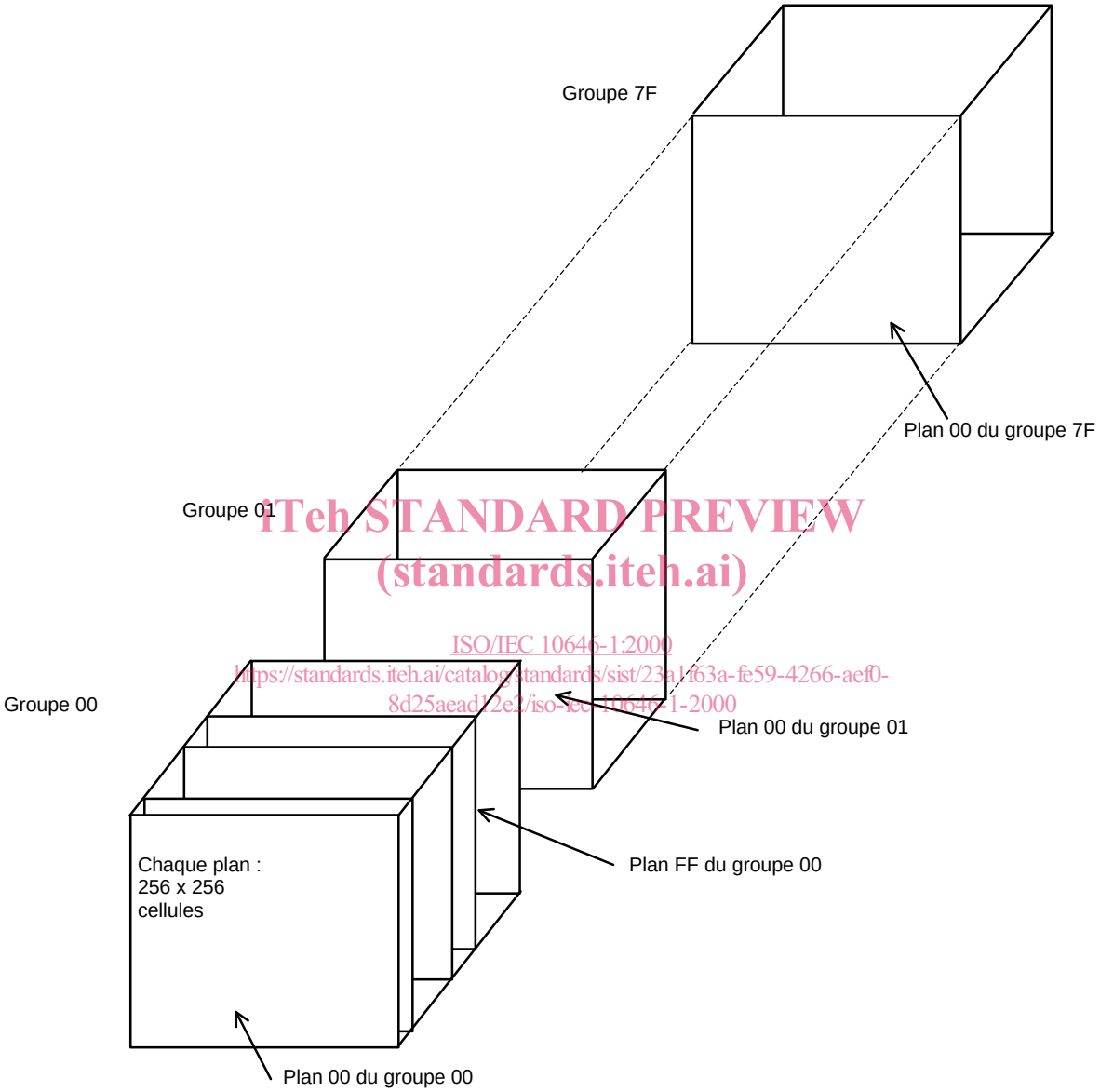
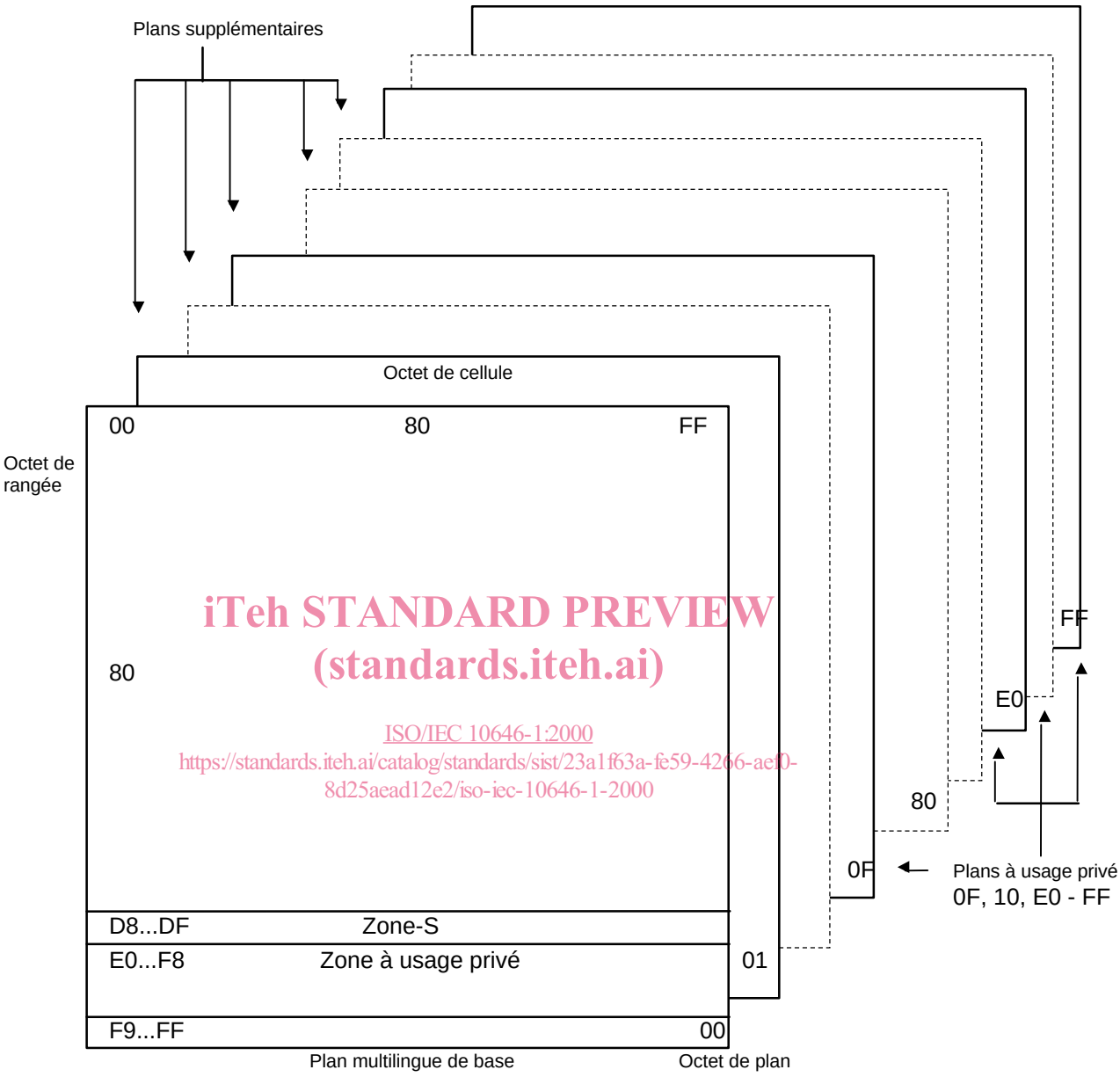


Figure 1 : Espace de codage complet du jeu universel de caractères codés sur plusieurs octets



NOTE : la Zone-S et la zone à usage privée sont décrites à l'article 8.

Figure 2 : Groupe 00 du jeu universel de caractères codés sur plusieurs octets

6.5 Identificateur abrégé des caractères

L'ISO/CEI 10646 définit un identificateur abrégé pour chaque caractère. Les identificateurs abrégés de deux caractères distincts sont distincts.

NOTE : Ces identificateurs abrégés sont indépendants de la langue dans laquelle la norme est écrite et sont conservés pour toute traduction de ce texte.

On définit les formes permises de notation d'un identificateur abrégé comme suit :

a. La forme à huit chiffres d'un identificateur abrégé sera composée de la suite des huit chiffres hexadécimaux qui représente la position de code du caractère (voir 6.2).

b. La forme à 4 chiffres d'un identificateur abrégé sera composée des 4 derniers chiffres de la forme à 8 chiffres. Cette forme n'est pas définie si les quatre premiers chiffres de la forme à 8 chiffres ne sont pas zéro, en d'autres termes, pour les caractères situés hors du plan multilingue de base.

c. Le caractère « -» (TIRET) peut, de manière optionnelle, précéder la forme à 8 chiffres de l'identificateur abrégé.

d. Le caractère « + » (SIGNE PLUS) peut, de manière optionnelle, précéder la forme à 4 chiffres de l'identificateur abrégé.

e. Le préfixe « U » (LETTRE MAJUSCULE LATINE U) peut, de manière optionnelle, précéder toutes les formes de l'identificateur abrégé définies dans les paragraphes a. à d. ci-dessus.

Les majuscules A à F, ainsi que U, qui peuvent apparaître au sein d'un identificateur abrégé peuvent être remplacés par leurs minuscules correspondantes.

La syntaxe complète de la notation des identificateurs abrégés, dans la forme de Backus-Naur, est donc :

{ U | u } [{+}xxxx | {-}xxxxxxxx]

où « x » représente un chiffre hexadécimal (0 à 9, A à F, ou a à f), par exemple:

-hhhhhhhh +kkkk
Uhhhhhhhh U+kkkk

où hhhhhhhh indique une forme à 8 chiffres et kkkk celle à 4 chiffres.

NOTE 1 : À titre d'exemple, l'identificateur abrégé de LETTRE MINUSCULE LATINE S LONG (voir les tables de la rangée 01 à l'article 26) peut prendre les formes suivantes :

0000017F	-0000017F	U0000017F	U-0000017F
017F	+017F	U017F	U+017F

Toutes les majuscules peuvent être remplacées par les minuscules correspondantes.

NOTE 2 : Il existe deux autres formes préfixées de la notation, pour lesquelles la lettre T (LETTRE MAJUSCULE LATINE T ou LETTRE MINUSCULE LATINE T) remplace la lettre U des formes préfixées correspondantes. Les formes de notation qui utilisent le préfixe T indiquent que l'identificateur abrégé fait référence à un caractère de la première édition de l'ISO/CEI 10646-1 (avant tout amendement), alors que les formes de notation qui utilisent le préfixe U indiquent toujours que l'identificateur abrégé se réfère au caractère de la version ISO/CEI 10646 la plus récemment amendée. Les identificateurs abrégés des formes T-xxxxxxxx et U-xxxxxxxx correspondants font référence au même caractère sauf quand xxxxxxxx est inclus dans l'intervalle fermé [00003400, 00004DFF]. Les formes de notation qui n'incluent pas de préfixe font toujours référence à la norme ISO/CEI 10646 comportant les corrections les plus récentes, sauf stipulation du contraire.

7 Caractéristiques particulières du JUC

Les caractéristiques suivantes s'appliquent au jeu de caractères codés dans sa totalité.

a) Les valeurs des octets P, R et C utilisés pour représenter des caractères graphiques doivent se situer dans la plage de 00 à FF. Les valeurs des octets G utilisés pour la représentation des caractères graphiques doivent se situer dans la plage de 00 à 7F. Les positions FFFE et FFFF ne doivent être utilisées dans aucun plan.

NOTE : La position de code FFFE est réservée à la « signature » (voir l'annexe H). La position de code FFFF peut être utilisée, notamment, pour des traitements internes qui requièrent une valeur numérique qui ne saurait être un caractère codé (par exemple pour signaler fin de tableau ou fin de texte). Puisqu'il s'agit de la plus grande valeur à deux octets, elle peut également être utilisée comme valeur finale dans un index de recherche binaire ou séquentielle.

b) À l'exception des positions réservées pour des caractères à usage privé ou pour des formats transformés, les positions auxquelles aucun caractère n'est attribué sont réservées pour une normalisation future et ne doivent pas être utilisées à d'autres fins. Les éditions à venir de l'ISO/CEI 10646 n'attribueront aucun caractère aux positions réservées aux caractères à usage privé ou aux formats transformés.

c) Le même caractère graphique ne doit pas être attribué à plus d'une position de code. Il existe des caractères graphiques ayant des formes semblables dans le jeu de caractères codés ; ils sont utilisés à des fins diverses et portent des noms de caractères différents.

8 Plan multilingue de base

Le plan 00 du groupe 00 constitue le plan multilingue de base (PMB). Le PMB peut être utilisé comme jeu de caractères codés à deux octets et sera alors appelé UCS-2 (voir 13.1).

Les positions 0000 0000 à 0000 001F du PMB sont réservées à des caractères de commande, la position 0000 007F étant réservée au caractère SUPPRESSION (voir article 15). Les positions 0000 0080 à 0000 009F sont réservées pour des caractères de commande.

Les positions 0000 D800 à 0000 DFFF sont réservées à l'utilisation de l'UTF-16 (voir annexe C). On nomme ces positions la zone-S.

Les positions 0000 E000 à 0000 F8FF sont réservées à l'usage privé (voir article 10). On nomme ces positions la zone à usage privé.

Les positions 0000 FFFE à 0000 FFFF sont réservées.

9 Autres plans

9.1 Plans réservés à la normalisation à venir

Les plans 11 à DF du groupe 00 et les plans 00 à FF des groupes 01 à 5F sont destinés à une future normalisation, ces positions ne doivent donc pas être utilisées à d'autres fins.

9.2 Plans accessibles par UTF-16

Chaque position des plans 01 à 10 du groupe 00 est reliée bijectivement à une suite de quatre octets selon la forme de représentation codée UTF-16 (voir annexe C). Cette forme est compatible avec la forme à deux octets du PMB UCS-2 (voir 13.1).

Il n'existe pas de correspondance entre la forme UTF-16 et les positions des plans 11 à FF du groupe 00 ou celles des plans 00 à FF pour les autres groupes.

10 Groupes, plans et zones à usage privé

10.1 Caractères à usage privé

L'ISO/CEI 10646 ne limite aucunement les caractères à usage privé. Ceux-ci peuvent être utilisés par l'utilisateur pour définir ses propres caractères. Il s'agit, par exemple, d'un besoin habituel pour les utilisateurs d'écritures idéographiques.

NOTE 1 : L'échange utile de caractères à usage privé nécessite un accord indépendant de l'ISO/CEI 10646 entre l'émetteur et le destinataire.

Les caractères à usage privé peuvent être utilisés pour des applications de caractères dynamiquement redéfinissables.

NOTE 2 : L'échange utile de caractères dynamiquement redéfinissables nécessite un accord indépendamment de l'ISO/CEI 10646 entre l'émetteur et le destinataire. L'ISO/CEI 10646 ne précise pas les techniques de définition ou de création des caractères dynamiquement redéfinissables.

10.2 Positions de code des caractères à usage privé

Les positions des 32 groupes, du groupe 60 au groupe 7Fsont réservés à l'usage privé.

Les positions du plan 0F, du plan 10 et des 32 plans de E0 à FF du groupe 00 sont à usage privé.

Les 6400 positions de code de E000 à F8FF du plan multilingue de base sont à usage privé.

Le contenu de ces positions n'est pas décrit dans l'ISO/CEI 10646 (voir 10.1).

11 Révision et mise à jour du JUC

La révision et la mise à jour du présent jeu de caractères codés seront effectuées par l'ISO/CEI JTC1/SC2.

NOTE : Dans les éditions à venir de l'ISO/CEI 10646, il est envisagé de conserver les noms et l'affectation des caractères de la présente édition.

12 Sous-ensembles

L'ISO/CEI 10646 définit des sous-ensembles de caractères graphiques codés utilisés lors d'un échange par des dispositifs de réception et d'émission.

Deux types de sous-ensembles peuvent être définis : les sous-ensembles limités et les sous-ensembles sélectionnés. Un sous-ensemble adopté peut comprendre l'un des deux ou une combinaison de ces deux types.

12.1 Sous-ensemble limité

Un sous-ensemble limité est composé d'une liste de caractères graphiques dans le sous-ensemble visé. Cette description permet l'interopérabilité, avec le présent jeu de caractères codés, d'applications et d'appareils qui utilisent d'autres codes.

Une déclaration de conformité concernant un sous-ensemble limité doit énumérer les caractères graphiques du sous-ensemble en donnant les noms des caractères graphiques ou les positions définis dans l'ISO/CEI 10646.

12.2 Sous-ensemble sélectionné

Un sous-ensemble sélectionné est composé d'une liste de collections de caractères graphiques définis dans l'ISO/CEI 10646. Les collections pouvant servir à la sélection sont énumérées à l'annexe A de chaque partie de l'ISO/CEI 10646. Un sous-ensemble sélectionné inclura d'office les cellules 20 à 7E de la rangée 00 du plan 00 du groupe 00.