Annexe C

Comparaison entre ISO/CEI 10646 et Unicode

Le Consortium Unicode a toujours collaboré étroitement avec le JTC1/SC2/GT2 de l'ISO/CEI, le groupe de travail responsable de l'élaboration de la norme internationale 10646. À l'heure actuelle, ces deux organismes se sont engagés à ce que le standard Unicode et la norme ISO 10646 demeurent synchronisés. Chacun de ces organismes utilise cependant son propre cadre de référence et, jusqu'à un certain point, sa propre terminologie. Cette annexe présente un bref historique des deux documents normatifs et les compare.

C.1 Historique

En 1984, le JTC1/SC2 de l'ISO/CEI définissait le mandat du groupe de travail numéro 2 (GT2) : « élaborer une norme établissant un répertoire de caractères graphiques des langues écrites du monde et son codage ».

Conscients des bénéfices d'une seule norme universelle de codage de caractères, les membres du Consortium Unicode collaborèrent avec les représentants de l'Organisation internationale de normalisation (ISO) pendant l'été et l'automne 1991. Les réunions entre les deux organismes permirent d'effectuer des changements acceptables par les deux parties à la version 1.0 d'Unicode et au premier projet de norme internationale de l'ISO/CEI, la PNI-1 10646. On attribua aux caractères du répertoire, désormais conjoint, un codage numérique unique. Cette collaboration déboucha sur la version 1.1 du standard Unicode.

En mai 1993, après plusieurs remaniements de texte afin de tenir compte des commentaires des membres votants, paraissait la norme internationale ISO/CEI 10646-1 : 1993, *Technologies de l'information — Jeu universel de caractères codés sur plusieurs octets (JUC) — Partie 1 : architecture et plan multilingue de base*. La version 1.1 du standard Unicode tenait compte des caractères supplémentaires introduits depuis le répertoire conjoint de la PNI-1 10646 ainsi que quelques remaniements de texte.

La fusion de la version 1.0 du standard Unicode et du PNI-1 10646 consistait à faire correspondre les valeurs numériques des caractères identiques et d'augmenter le répertoire conjoint de quelques groupes de caractères présents dans le PNI-1 10646 mais absents du standard Unicode. C'est pourquoi les valeurs de code des caractères de l'ISO/CEI 10646-1 : 1993 UCS-2 et de la version 1.1 du standard Unicode sont rigoureusement identiques. Les versions 2.0 et 3.0 virent l'ajout de nouveaux caractères afin de les faire correspondre aux additions récentes de l'ISO/CEI 10646-1 : 2000. Le tableau C1 fournit ci-dessous la chronologie de cette collaboration.

Tableau C-1. Historique

Année	Version	Description
1984		Création du GT2 du JTC1/SC2
1989	PP 10646	Proposition préliminaire de l'ISO, indépendante d'Unicode
1990	Prépublication Unicode	Prépublication pour examen du projet Unicode
1990	PNI-1 10646	Premier projet, indépendant d'Unicode
1991	Unicode 1.0	Publication
1992	Unicode 1.0.1	Modifié pour des raisons de compatibilité avec la 10646
1992	PNI-2 10646	Deuxième projet, fusionné avec Unicode
1993	NI 10646-1 : 1993	Norme fusionnée (uniquement en anglais)
1993	Unicode 1.1	Révisé pour correspondre à la NI 10646-1 : 1993
1995	Amendements à la 10646	Réalignement du coréen, suppléments
1996	Unicode 2.0	Synchronisé avec les amendements de l'ISO 10646
1998	Unicode 2.1	Ajout du signe euro et quelques corrections
1999	Amendements à la 10646	Suppléments
2000	Unicode 3.0	Synchronisé avec la 2e édition de l'ISO 10646
2000	NI 10646-1 : 2000	10646, partie 1, 2e édition, publication en français et en anglais reprenant les amendements publiés jusqu'alors
2001	Unicode 3.1	Synchronisé avec le ISO 10646-2 à deux caractères près : U+03F4 et U+03F5.
2001	NI 10646-2 : 2001	10646, partie 2, 1e édition.

Unicode 1.0

Le répertoire conjoint publié dans l'ISO/CEI 10646 est un sur-ensemble du répertoire de la version 1.0 du standard Unicode tel qu'amendé par la version 1.0.1 du même standard. La version 1.0 du standard Unicode avait été corrigée par l'additif intitulé Unicode 1.0.1 afin d'en faire un sous-ensemble strict de l'ISO/CEI 10646. Cela donna lieu à la fois au déplacement et à l'élimination d'un petit nombre de caractères.

Unicode 2.0

La version 2.0 du standard Unicode reprenait le répertoire de la version 1.1 du standard Unicode (et de la NI 10646), ainsi que les sept amendements suivants apportés à la NI 10646 :

Amendement 1	UTF-16
Amendement 2	UTF-8
Amendement 3	Codage des commandes C1
Amendement 4	Suppression de l'annexe G : UTF-1
Amendement 5	Collection des caractères coréens hangûl
Amendement 6	Collection des caractères tibétains
Amendement 7	33 caractères supplémentaires (hébreu, eszett, dong)

En outre, la version 2.0 du standard Unicode comprenait également le rectificatif technique n° 1 (portant sur le nouveau nom de AE qui, de ligature, devenait une lettre) et des rectificatifs au texte de l'ISO/CEI 10646 qui s'appliquait également au standard Unicode. Le symbole euro et le caractère de remplacement d'objet seront ajoutés lors de la version 2.1, suivant l'amendement 18 de l'ISO/CEI 10646-1.

Unicode 3.0

La version 3.0 du standard Unicode cadre avec la deuxième édition de l'ISO/CEI 10646-1. Cette dernière reprend tous les amendements apportés à la 10646-1 publiés jusqu'alors. Cette liste comprend les sept premiers amendements ainsi que les suivants :

Amendement 8	Addition de l'annexe T : Procédure pour l'unification et la disposition des idéogrammes CJC
Amendement 9	Identificateurs des caractères
Amendement 10	Collection des caractères éthiopiens
Amendement 11	Collection des syllabaires autochtones canadiens
Amendement 12	Collection des caractères chérokîs
Amendement 13	Idéophonogrammes unifiés CJC issus de sources supplémentaires (extension horizontale)
Amendement 14	Collection des caractères syllabiques yi et des clés yi
Amendement 15	Collection des caractères regroupant les clés K'ang-hsi et des nombres Hang-tcheou
Amendement 16	Collection des caractères regroupant les combinaisons Braille
Amendement 17	Supplément A aux idéophonogrammes unifiés CJC (extension verticale)
Amendement 18	Collection de caractères regroupement des lettres et caractères divers (parmi lesquels le symbole monétaire euro)
Amendement 19	Collection des caractères runiques
Amendement 20	Collection des caractères ogamiques
Amendement 21	Collection des caractères singhalais
Amendement 22	Collection des caractères symboliques pour le clavier
Amendement 23	Collection des caractères supplémentaires bopomofo ainsi que quelques autres caractères
Amendement 24	Collection des caractères thâna
Amendement 25	Collection des caractères khmers
Amendement 26	Collection des caractères birmans
Amendement 27	Collection des caractères syriaques
Amendement 28	Caractères de description idéophonographique
Amendement 29	Mongol
Amendement 30	Supplément latin et autres caractères
Amendement 31	Supplément tibétain

La seconde édition de la 10646-1 comprend également le contenu du correctif technique n°2 ainsi que tous les correctifs apportés au texte jusqu'alors.

La synchronisation de la version 3.0 du standard Unicode et de la deuxième édition de l'ISO/CEI 10646-1 signifie que le répertoire, le codage et le nom[1] des caractères sont identiques dans les deux documents et que tous les autres amendements apportés à la 10646-1 et qui ont une incidence sur le texte du standard Unicode ont été pris en compte lors de cette révision du standard Unicode.

La version 3.1 introduit 44.946 nouveaux caractères codés. Parmi les nouveaux caractères, on retrouve l'italique, le gotique, des symboles mathématiques et près de 43.000 nouveaux idéogrammes CJC. C'est la première version qui inclut des caractères codés au-delà de l'espace de codage à 16 bits original (ou encore au-delà du plan zéro encore appelé PMB).

C.2 Formes de codage de l'ISO/CEI 10646

L'ISO/CEI 10646 définit deux formes de codage concurrentes :

- Un codage sur quatre octets (32 bits) qui comprend 2^{31} positions de code. Ces positions de code sont conceptuellement divisées en 128 *groupes* de 256 *plans*, chacun de ces plans étant constitué de 256 *rangées* de 256 *cellules*.

- Un codage sur deux octets, ou un seizet (16 bits), qui comprend le plan zéro, le plan multilingue de base (PMB).

La forme sur 32 bits est connue sous le nom d'UCS-4 (jeu universel de caractères sur 4 octets) ; à la forme sur 16 bits elle porte le nom d'UCS-2 (jeu universel de caractères sur 2 octets).

Les numéros de code de 0 à 65 535 décimal (0–FFFF hexadécimal) peuvent être représentés à l'aide d'une valeur de code de caractère à 16 bits. Les caractères essentiels (c'est-à-dire les caractères des normes mondiales principales) sont affectés au PMB. L'ISO/CEI 10646 ne définit pas pour l'instant de caractères dans d'autres plans. L'ISO/CEI 10646-2 définira le codage des caractères affectés aux autres plans.

C.3 Formats de transformation du JUC

UTF-8

Le terme UTF-8 signifie *format transformé du JUC sur 8 bits*. UTF-8 est un des formats de représentation codée de tous les caractères de l'ISO/CEI 10646. Le format défini par l'ISO/CEI est identique à celui décrit pour UTF-8 à la section *2.3, Formes de codage*.

UTF-8 peut être utilisé dans la transmission de texte de données au travers de systèmes de communication qui supposent que les valeurs d'octets situées entre x00 et x7F suivent la définition de l'ISO/CEI 4873 et comprennent un ensemble C0 de fonctions de commande respectant la structure à 8 bits de l'ISO/CEI 2022. L'UTF-8 évite également d'utiliser les

[1] Il n'existe pas de noms officiels en français pour le standard Unicode. Nous avons donc utilisé les noms de caractères officiels de la version française de l'ISO/CEI 10646.

valeurs d'octets dans cet intervalle qui ont un sens particulier lors de l'analyse des noms de fichiers dans la plupart des systèmes d'exploitation.

L'amendement 2 de l'ISO/CEI 10646 définit pour la première fois l'UTF-8 ; il fait partie de la deuxième édition de l'ISO/CEI 10646-1.

La définition d'UTF-8 à l'annexe D de l'ISO/CEI 10646-1 :2000 permet également l'utilisation de suites de cinq ou six octets dans le but de coder les caractères n'appartenant pas à l'intervalle des caractères Unicode. Ces suites de cinq ou six octets sont illégales en UTF-8 quand celui-ci est utilisé comme transformation de caractères Unicode. L'ISO/CEI 10646 ne permet pas l'utilisation de seizet d'indirection non appariés, ni les valeurs U+FFFE et U+FFFF ; cependant elle permet l'utilisation de non-caractères.

UTF-16

Le terme UTF-16 signifie *format transformé du JUC sur 16 bits*. UTF-16 est le codage de l'ISO/CEI équivalent au standard Unicode ; il comprend l'utilisation de seizets d'indirection, voir le chapitre *3, Conformité.* En UTF-16, chaque valeur de code UCS-2 représente sa propre valeur. On représente les valeurs de code de l'ISO/CEI 10646 appartenant aux plans 1 à 16 à l'aide de codes spéciaux, appelés seizets d'indirection. UTF-16 définit la transformation entre les codes de position UCS-4 des plans 1 à 16 du groupe 00 et les paires de codes spéciaux. Cette transformation est identique à celle définie dans le standard Unicode au paragraphe *D.28* à la section *3.7, Seizets d'indirection.* On trouvera sur le disque qui accompagne ce livre des exemples de code permettant de transformer de l'UCS-4 en caractères Unicode avec seizets d'indirection.

UTF-16 permet de représenter le PMB et les 16 plans suivants. Cette restriction n'est pas injustifiée car le JTC1/SC2/GT2 de l'ISO n'a aucune intention d'affecter des caractères au-delà du plan 14 car cela remettrait en cause sa synchronisation avec le standard Unicode. Les plans 15 et 16 ($000F0000..000FFFFF_{16}$ et $00100000..0010FFFF_{16}$) sont à usage privé. Les valeurs de codes UCS-4 pour les zones à usage privé des groupes 60 à 7F et celles des plans E0 à FF du groupe 00 ne sont pas accessibles par UTF-16. On décourage fortement l'utilisation de codes à utilisation privée car les données codées à l'aide de ces valeurs de code ne pourront être échangées avec des applications qui mettent en œuvre Unicode 3.1. À leur place, il faut plutôt utiliser les plans 15 et 16.

Les applications qui échangent des données ISO/CEI 10646 dont certaines proviennent des plans 1 à 16 devraient utiliser l'UTF-16 comme forme de stockage implicite en absence d'information à l'effet du contraire.

UTF-32

Cette transformation ne fait partie de la norme Unicode 3.1 (cf. l'annexe normative d'Unicode n° 19). Comme on l'a vu, l'ISO/CEI 10646 définit un format de codage sur 4 octets dénommé UCS-4. Étant donné qu'UTF-32 est tout simplement un sous-ensemble des caractères d'UCS-4, il se conforme donc aussi bien à l'ISO/CEI 10646 qu'au standard Unicode.

Le terme UTF-32 est un terme synonyme à UCS-4, avec l'exigence complémentaire de respect de la sémantique Unicode.

C.4 Sous-ensembles de l'ISO/CEI 10646

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.

Un sous-ensemble limité est composé d'une liste de caractères graphiques dans le sous-ensemble visé. 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 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, en d'autres mots la collection « latin de base ». La majorité des collections définies par la 10646 correspond à un seul bloc de caractères, il existe cependant quelques collections qui regroupent ou divisent des blocs, comme les « séparateurs de formatage » ou les « caractères combinatoires ».

C.5 Niveaux de mise en œuvre de l'ISO/CEI 10646

L'ISO/CEI 10646 définit trois niveaux de mise en œuvre :

- Au niveau 1 de mise en œuvre, un flux de données ISO/CEI 10646 ne doit contenir ni représentations codées de caractères combinatoires ni caractères du bloc JAMOS HANGÛL. Est également exclue une liste de signes diacritiques et combinatoires reprise à l'annexe B.1 de la norme.

- Le niveau 2 de mise en œuvre exclut l'utilisation de représentations codées des caractères diacritiques ou combinatoires énumérés à l'article B.2 de la norme. Cette liste est un sous-ensemble de la liste de l'article B.1.

- Au niveau 3 de mise en œuvre, un flux de données peut contenir des représentations codées de tous les caractères.

C.6 Identification des fonctionnalités de l'ISO/CEI 10646

Dans son optique axée sur le transport entre un dispositif émetteur et récepteur, l'ISO/CEI 10646 définit des séquences d'échappement compatibles avec l'ISO/CEI 2022 qui permettent d'identifier les caractéristiques dont un dispositif est doté. Ces séquences de désignation permettent de préciser une forme de représentation codée du JUC et du niveau de mise en œuvre défini. Voici quelques exemples de séquences de désignation :

ESC 02/05 02/15 04/00
    UCS-2 et mise en œuvre de niveau 1
ESC 02/05 02/15 04/01
    UCS-4 et mise en œuvre de niveau 1
ESC 02/05 02/15 04/03
    UCS-2 et mise en œuvre de niveau 2
ESC 02/05 02/15 04/04
    UCS-4 et mise en œuvre de niveau 2
ESC 02/05 02/15 04/05

UCS-2 avec mise en œuvre de niveau 3
ESC 02/05 02/15 04/06
UCS-4 avec mise en œuvre de niveau 3.

Il existe également des séquences de désignation pour les différents formats de transformation et pour indiquer le passage d'une suite de caractères conforme à l'ISO/CEI 10646 à une suite de caractères conforme à l'ISO/CEI 2022.

Dans tous les cas de figure, les séquences d'échappement sont définies comme étant des octets étendus par des zéros afin de faire correspondre leur taille avec celle des caractères dans le codage courant. Ainsi, le caractère ÉCHAPPEMENT, transcrit ESC ci-dessus, est-il égal à l'octet 1B. Quand il fait partie d'un flux de caractères UCS-2, on le représente par 00 1B et quand il fait partie d'un flux de caractères UCS-4, il est alors représenté par 00 00 00 1B.

C.7 Plans réservés et à usage privé de l'ISO/CEI 10646

Les positions des 32 groupes, situés du groupe 60 au groupe 7F, sont 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 6 400 positions de code de E000 à F8FF du plan multilingue de base sont à également usage privé. Le contenu de ces positions n'est pas décrit dans l'ISO/CEI 10646.

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.

Chaque position des plans 01 à 10 du groupe 00 est reliée bijectivement à une suite de quatre octets en format UTF-16. 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. Cf. également UTF-32 à la section *C.3, Formats de transformation JUC.*

C.8 Zones du PMB

Nom de la zone	Début	Fin	Nombre de positions	Description
A	0000 0000	0000 4DFF	19.903	Alphabets, symboles, divers CJC, hangûl, idéogrammes. Il semble manquer 65 caractères à cette zone, cependant les caractères de commande C0 et C1 ainsi que le caractère 0000 007F sont réservés ; ils ne sont donc pas comptabilisés au sein de la zone A.
I	0000 4E00	0000 9FFF	20.992	Idéophonogrammes chinois, japonais et coréens unifiés.
O	0000 A000	0000 D7FF	14.336	Hangûl.
S	0000 D800	0000 DFFF	2.048	Zone d'indirection, cf. UTF-16.
R	0000 E000	0000 FFFD	8.190	Caractères à usage privé, formes de présentation d'autres caractères du répertoire et les caractères de compatibilité.

Nom de la zone	Début	Fin	Nombre de positions	Description
				0000 FFFE et 0000 FFFF ne sont pas comptabilisés puisque les seizets FFFE et FFFF ne sont utilisés par aucun plan. FFFE sert de signature de codage. On est également assuré que FFFF n'est pas un caractère. FFFF peut donc servir à indiquer la fin d'une chaîne de caractères.

C.9 Identification des fonctionnalités pour le standard Unicode

L'ISO/CEI 10646 fournit un mécanisme qui permet d'indiquer un certain nombre de paramètres de mise en œuvre, lesquels permettent d'engendrer ce qu'on pourrait appeler des instances de la norme. L'ISO/CEI 10646 est cependant dépourvue de mécanisme pour décrire une mise en œuvre nommée « Unicode ». Dans son intégralité, toutefois, on peut considérer que le standard Unicode reprend tout le répertoire de l'ISO/CEI 10646 et qu'il a les fonctionnalités de l'ISO 10646 suivantes :

- un sous-ensemble sélectionné composé de la collection n° 300 (le PMB);
- UTF-16 (si les seizets d'indirection sont utilisés, autrement UCS-2);
- niveau de mise en œuvre n° 3 (signes combinatoires et caractères précomposés permis);
- dispositif de type n° 1 (dispositif de réception avec pleine capacité de retransmission).

À ces fonctionnalités, il convient d'ajouter qu'Unicode codifie une sémantique que la 10646 ne précise pas.

Peu d'applications utiliseront sans doute tous les caractères définis dans le plan multilingue de base de l'ISO/CEI 10646. Les clauses de conformité des deux documents abordent cette situation sous deux angles très différents. L'ISO/CEI fournit, comme on l'a vu, un mécanisme qui permet de préciser les sous-ensembles du répertoire adoptés; de la sorte les mises en œuvre peuvent simplement ignorer les caractères qui ne sont pas inclus (cf. l'annexe normative A de l'ISO/CEI 10646). Par contre, une mise en œuvre de la norme Unicode devra traiter tous les caractères à un niveau minimal : elle devra pouvoir les stocker et les retransmettre intacts. Le standard Unicode englobe tout le plan multilingue de base de l'ISO/CEI 10646 sans nécessiter qu'un sous-ensemble soit mis en œuvre.

Le standard Unicode ne permet pas d'indiquer qu'un flux d'octets est constitué de caractères Unicode bien que cette fonction puisse en partie être remplie par l'indicateur d'ordre des octets (U+FEFF ESPACE INSÉCABLE SANS CHASSE). L'ISO/CEI, par contre, spécifie des séquences d'échappement de l'ISO/CEI 2022 qui permettent l'identification d'une forme de représentation codée du JUC et du niveau de mise en œuvre. L'ISO/CEI 10646 permet également d'utiliser la « signature » U+FEFF. Il convient de signaler cette convention de signature facultative qui permet de distinguer entre les formes UCS-2 et UCS-4. Cette méthode est résumée à la section *14.6, Codes spéciaux*.

C.10 Noms des caractères

Les noms de caractères anglais d'Unicode respectent les lignes directrices de l'ISO précisées à l'annexe K de l'ISO/CEI 10646. Les noms français repris dans cet ouvrage correspondent également à ceux de la version officielle française de l'ISO/CEI 10646 et respectent donc, *ipso facto*, les règles d'affectation de nom.

Dans la conception de l'ISO/CEI, l'unicité du nom de caractère permet à la fois d'affecter une sémantique aux caractères et d'assurer une correspondance entre différentes normes. Pour Unicode, la sémantique des caractères est fournie par des noms optionnels (les *alias*), les annotations, les propriétés de caractères et les spécifications fonctionnelles mentionnées au chapitre *3, Conformité*. Quant à la correspondance avec d'autres normes elle est assurée à l'aide de tables explicites.

C.11 Spécifications fonctionnelle des caractères

La base d'une norme de codage de caractères est une correspondance entre des valeurs de code et des caractères. Parfois, cependant, la sémantique ou l'identité d'un caractère peut demeurer ambiguë. Il est certain qu'un caractère ne se confond pas avec le glyphe représentatif qui l'illustre dans ce livre. C'est pourquoi Unicode fournit les renseignements sémantiques nécessaires pour décrire les caractères qu'il code.

Le standard Unicode comporte donc bien plus qu'un tableau de valeurs de code, contrairement à l'ISO/CEI 10646. Il comprend également un ensemble complet de spécifications fonctionnelles pour les caractères et les données qu'il code, ainsi qu'une imposante somme de renseignements linguistiques ou typographiques destinés à permettre aux personnes chargées de la mise en application de mieux comprendre la manière dont les caractères se comportent. Le standard Unicode spécifie propriétés et algorithmes. Les mises en œuvres conformes du standard Unicode se conforment également à l'ISO/CEI 10646, niveau de mise en œuvre 3.

Les mises en œuvre conformes à l'ISO/CEI 10646 peuvent être conformes au standard Unicode pour autant qu'elles se conforment aux spécifications supplémentaires qui s'appliquent aux caractères de leurs sous-ensembles adoptés et qu'elles prennent en charge tous les caractères n'appartenant pas aux sous-ensembles adoptés de la façon indiquée à la section *C.9, Identification des fonctionnalités pour le standard Unicode.*