# Chapitre 1

# Introduction

Le standard Unicode[1] est un mécanisme universel de codage de caractères. Il définit une manière cohérente de coder des textes multilingues, facilite l'échange de données textuelles à l'échelle planétaire et crée ainsi les prémisses pour tout logiciel international. En tant que codage implicite de HTML et XML, le standard Unicode constitue un socle ferme pour l'Internet et les nouvelles méthodes d'affaire d'un monde de plus en plus réticulé. Obligatoire pour la plupart des nouveaux protocoles de l'Internet, mis en œuvre dans tous les systèmes d'exploitation et langages informatiques modernes (p.ex. Java), Unicode est la base de tout logiciel qui doit fonctionner aux quatre coins du monde.

Grâce à Unicode, l'industrie informatique assure la stabilité de ses données en évitant la prolifération pléthorique des jeux de caractères, augmente l'interopérabilité et l'échange de données au niveau mondial, enfin Unicode simplifie le développement de logiciels tout en en réduisant les coûts.

Bien qu'Unicode ait été conçu sur le modèle du jeu de caractères ASCII, il va bien au-delà des capacités rudimentaires d'ASCII qui ne code que les lettres majuscules et minuscules A à Z. Unicode permet de coder tous les caractères utilisés par toutes les langues écrites du monde, plus d'un million de caractères sont réservés à cet effet. Tous les caractères, quelle que soit la langue dans laquelle ils sont utilisés, sont accessibles sans aucune séquence d'échappement. Le codage de caractère Unicode traite les caractères alphabétiques, les caractères idéographiques et les symboles de manière équivalente, avec comme conséquence que tous ces caractères peuvent se côtoyer dans n'importe quel ordre avec une égale facilité (cf. *Figure 1-1*).

Le standard Unicode attribue à chacun de ses caractères une valeur numérique et un nom. À ce chapitre, il ne diffère guère des autres standards ou normes de codage de caractères depuis l'ASCII. Cependant Unicode fournit d'autres renseignements cruciaux afin de s'assurer que le texte produit soit lisible : la casse de caractère, sa directionalité et ses propriétés alphabétiques. Unicode définit également des renseignements sémantiques et comprend des données d'applications, tels que des tableaux de correspondance de casse ou des conversions entre les répertoires de jeux de caractères internationaux, nationaux ou encore utilisés par l'industrie. Le consortium Unicode fournit ces renseignements complémentaires afin d'assurer le développement de mises en œuvres cohérentes et l'échange de données Unicode.

À l'heure actuelle, les données Unicode peuvent être codées sous deux formes : une forme implicite de 16 bits et une forme de 8 bits dénommée UTF-8 conçue pour faciliter son utilisation sur les systèmes ASCII préexistants. La version 3.0[2] du standard Unicode est identique à la norme internationale ISO/CEI 10646 en ce qui a trait à l'affectation des

---

[1] Généralement toute mention au « standard Unicode » s'applique également à la norme internationale ISO/CEI 10646. Pour une comparaison détaillée entre ces deux textes, voir l'annexe *C, Comparaison entre ISO/CEI 10646 et Unicode*.

[2] La version 3.1 d'Unicode correspond à deux caractères près à l'ISO/CEI 10646-1:2000 et ISO/CEI 10646-2. Il s'agit de U+03F4 et U+03F5, ils feront partie du prochain amendement n°1 de l'ISO 10646-1:2000, ils sont inclus dans Unicode 3.1 afin de permettre des décompositions cohérentes aux symboles alphanumériques mathématiques

caractères et à leurs noms. Toute application qui se conforme à Unicode se conforme donc à l'ISO/CEI 10646.

L'utilisation d'un codage à 16 bits signifie qu'il existe des valeurs de code pour plus de 65.000 caractères. Bien que ce nombre soit insuffisant pour coder directement tous les caractères utilisés dans les langues principales du monde, le standard Unicode et l'ISO/CEI 10646 définissent un mécanisme d'extension appelé UTF-16 (dont les caractères d'extension sont appelés *seizets d'indirection* [3]). Celui-ci permet de coder plus d'un million de caractères supplémentaires sans devoir utiliser de codes d'échappement. Ce nombre suffit à coder tous les caractères connus, y compris ceux utilisés par toutes les écritures historiques de la planète.

**Figure 1-1. ASCII large**

| Texte ASCII/8859-1 | |
|---|---|
| T | 0101 0100 |
| e | 0110 0101 |
| x | 0111 1000 |
| t | 0111 0100 |
| e | 0110 0101 |
| | 0010 0000 |
| A | 0100 0001 |
| S | 0101 0011 |
| C | 0100 0011 |
| I | 0100 1001 |
| I | 0100 1001 |
| / | 0010 1111 |
| 8 | 0011 1000 |
| 8 | 0011 1000 |
| 5 | 0011 0101 |
| 9 | 0011 1001 |
| - | 0010 1101 |
| 1 | 0011 0001 |

| Texte Unicode/ISO 10646 | |
|---|---|
| T | 0000 0000 0101 0100 |
| e | 0000 0000 0110 0101 |
| x | 0000 0000 0111 1000 |
| t | 0000 0000 0111 0100 |
| e | 0000 0000 0110 0101 |
| | 0000 0000 0010 0000 |
| 天 | 0101 1001 0010 1001 |
| 地 | 0101 0111 0011 0000 |
| | 0000 0000 0010 0000 |
| س | 0000 0110 0011 0011 |
| ل | 0000 0110 0100 0100 |
| ا | 0000 0110 0010 0111 |
| م | 0000 0110 0100 0101 |
| | 0000 0000 0010 0000 |
| ᾷ | 0000 1111 0000 0110 |
| α | 0000 0011 1011 0001 |
| ⋦ | 0010 0010 0111 0000 |
| γ | 0000 0011 1011 0011 |

[3] Depuis l'approbation par le consortium Unicode du format UTF-32 (cf. le rapport technique d'Unicode, n° 19), il existe désormais trois formes de codage. Cette dernière forme (ainsi que la forme UCS-4 de l'ISO/CEI 10646 dont elle s'inspire) permet d'adresser directement plus d'un million de caractères.

# 1.1 Domaine d'application

La version 3.1 du standard Unicode comprend 94.140 caractères issus de tous les systèmes d'écriture du monde. Cet ensemble est plus que suffisant pour satisfaire les besoins de communication moderne mais il permet également de coder la plupart des formes classiques de nombreuses langues. Parmi ces écritures, on retrouve les écritures alphabétiques européennes, les écritures de droite à gauche du Moyen-Orient et les écritures de l'Asie. Le jeu unifié han compte 71.039[4] caractères idéographiques définis par des normes nationales ou industrielles de Chine, du Japon, de Corée, de Taïwan, du Viêt-Nam et de Singapour. Le standard contient également de nombreux signes de ponctuation, des symboles mathématiques et techniques, des formes géométriques et des dingbats ou caractères de casseau.

De nombreuses nouvelles écritures ont fait leur apparition dans les versions 3.0 et 3.1, parmi celles-ci : l'éthiopien, le syllabaire autochtone canadien, le chérokî, le singhalais, le syriaque, le birman, le khmer, le mongol, le braille. l'ancien italique, le gotique et des idéogrammes supplémentaires. On trouve au chapitre *2, Structure générale,* une vue d'ensemble de l'attribution des caractères.

Remarquons, cependant, qu'Unicode ne code pas les caractères personnels, neufs, rarement échangés ou d'usage privé pas plus que les logos ou les graphiques. Les systèmes graphiques sans rapport avec le texte, comme les notations chorégraphiques, ne relève pas du standard Unicode. Les variantes de police sont explicitement exclues de tout codage. Six mille quatre cents valeurs de code sont réservées dans le codage de base à 16 bits à la zone à usage privé. Cette zone peut servir à coder des caractères qui ne font pas partie du répertoire Unicode.

Il existe 7.793 valeurs de code non affectées dans le *plan multilingue de base* (PMB), elles pourraient l'être à la suite d'une normalisation ultérieure. Il existe également 872.532 points de code non affectés dans les plans complémentaires. Les plans complémentaires contiennent 131.068 points de code à usage privé qui s'ajoutent aux 6.400 prévus dans le plan multilingue de base.

## Domaine d'application normatif

Le standard Unicode est un sur-ensemble de tous jeux de caractères couramment utilisés aujourd'hui. Il comprend les caractères des principales normes nationales et internationales ainsi que les plus importants jeux de caractères industriels. Ainsi, Unicode incorpore l'ISO/CEI 6937 et la famille des normes ISO/CEI 8859, la norme SGML ISO/CEI 8879 et des normes bibliographiques telles que l'ISO 5426. Parmi les normes nationales importantes incluses dans Unicode, on retrouve les normes ANSI Z39.64, KS C 5601, JIS X 0209, GB 2312 et CNS 11643. Les pages de codes de l'industrie informatique ne font pas défaut, on en retrouve provenant d'Adobe, d'Apple, de Fujitsu, de Hewlett-Packard, d'IBM, de Lotus, de Microsoft, de NEC et Xerox.

Pour une liste complète des normes nationales et internationales (ISO) utilisées comme sources, veuillez vous référer à l'annexe *B, Sources bibliographiques*.

[4] Dont 70.207 idéogrammes han unifiés et 832 idéogrammes CJC de compatibilité.

### Nouveaux caractères

Le standard Unicode continue de répondre aux besoins d'une industrie jeune et en pleine évolution et intègre constamment de nouveaux caractères importants. Ainsi, lorsqu'il devint nécessaire de coder le signe monétaire de l'euro, le consortium publia la version 2.1 du Standard Unicode en y attribuant une valeur à ce signe.

En tant que mécanisme de codage universel de caractères, le standard Unicode se doit également de suppléer aux besoins des philologues. Afin de préserver l'héritage culturel mondial, il est également important que les principales écritures archaïques soient codées au fur et à mesure que des propositions de codage voient le jour.

Pour proposer de nouveaux caractères, veuillez contacter votre représentant national auprès de l'ISO < http://www.iso.ch/members/index.html?%138FR>.

## 1.2 Principes de conception

Historiquement, le but principal du standard Unicode fut de remédier à deux écueils sérieux et fréquents dans la plupart des programmes informatiques multilingues. Le premier de ces problèmes était l'ambiguïté du codage des caractères au sein des polices. Ainsi, les octets 0x00 à 0xFF sont souvent utilisés pour coder des caractères et des symboles totalement différents. Le second problème important consistait dans l'utilisation de multiples codes de caractère incompatibles provenant de normes nationales et industrielles contradictoires. Dans les environnements informatiques d'Europe occidentale, on trouve ainsi des conflits entre la page de code Windows 1252 (Latin 1) et l'ISO/CEI 8859-1. Dans le cas de logiciels qui prennent en charge les idéogrammes d'Extrême-Orient, le même ensemble d'octets utilisés en ASCII peut également servir de deuxième octet pour les caractères à deux octets. Dans ces cas-là, les logiciels doivent être capables de distinguer les caractères ASCII des caractères à deux octets.

L'espace de code à 7 bits de l'ASCII ou celui à 8 bits de ses successeurs, bien qu'ils soient quasi omniprésents, n'offrent que 128 et 256 positions de code respectivement. Ces espaces de code sont inefficaces et totalement inadéquats pour des environnement informatiques internationaux.

À la naissance en 1988 du projet Unicode, les utilisateurs les plus défavorisés par ce manque de norme internationale cohérente regroupaient les éditeurs de logiciels scientifiques et mathématiques, les éditeurs de livres, les services de renseignements bibliographiques et les chercheurs universitaires. Récemment, l'industrie informatique a adopté une perspective résolument plus internationale et développe des logiciels qui peuvent facilement être adaptés aux besoins linguistiques et culturels de nombreux nouveaux marchés. La croissance fulgurante de l'Internet a simplement accru le besoin d'un jeu de caractère normalisé susceptible d'être utilisé à l'échelle planétaire.

Les concepteurs du standard Unicode envisageaient une méthode uniforme d'identification des caractères plus efficace et plus souple que les systèmes de codage précédents. Le nouveau système devait répondre aux besoins techniques et multilingues tout en codant un grand éventail de caractères qui permettent d'assurer les besoins typographiques de qualité professionnelle et de la micro-édition dans le monde entier.

Le standard Unicode fut donc conçu pour être :

- *Universel.* Le répertoire doit être suffisamment étendu pour comprendre tous les caractères susceptibles d'être utilisés dans les échanges de textes habituels, y compris les principaux jeux de caractères internationaux, nationaux ou industriels.

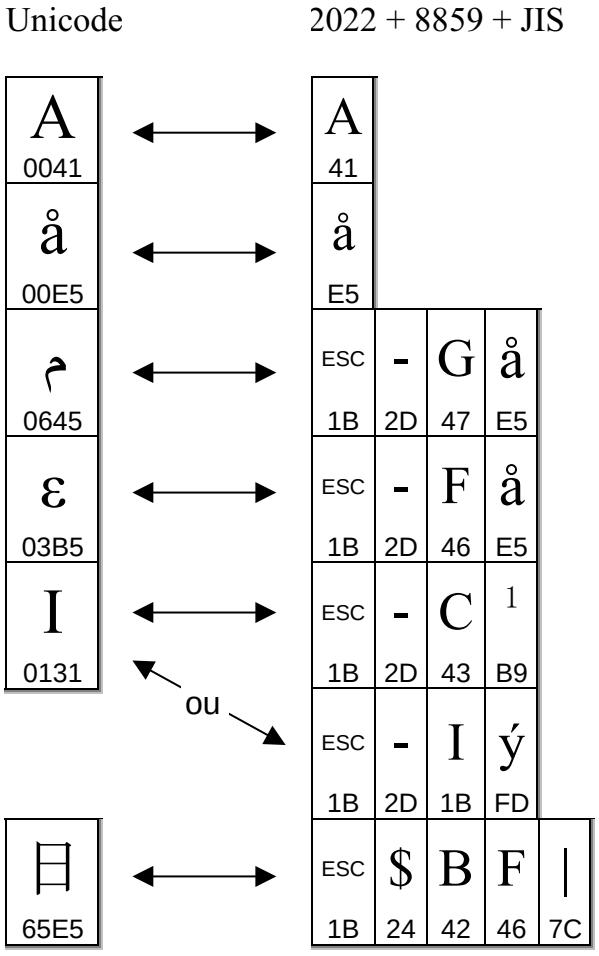
- *Efficace.* Le texte brut doit être facile à analyser : les logiciels ne doivent pas maintenir une variable d'état ou rechercher des séquences d'échappement, la synchronisation de caractère à partir de n'importe quel point dans le flux de caractères doit être rapide et non ambigu.

- *Uniforme.* Un jeu de caractères de largeur fixe permet de trier, de repérer, d'afficher et d'éditer des textes efficacement.

- *Non ambigu.* Toute valeur de 16 bits représente toujours le même caractère.

La figure 1-2 illustre quelques-unes de ces caractéristiques en opposant le codage Unicode à une série de jeux de caractères à un octet qui utilisent des séquences d'échappement pour modifier le sens des octets.

**Figure 1-2. Universel, efficace et non ambigu**



# 1.3 Traitement textuel

Traiter du texte implique deux processus : le traitement proprement dit et le codage. Quand l'utilisateur d'un logiciel de traitement de texte saisit du texte au moyen de son clavier, le système d'exploitation de l'ordinateur reçoit un message indiquant que l'utilisateur a enfoncé la combinaison de touches correspondant, par exemple, au « T », caractère qu'il code avec la valeur U+0054. Le texteur stocke le chiffre en mémoire et le passe au logiciel responsable d'afficher le caractère à l'écran. Ce logiciel de rendu, qui peut faire partie du gestionnaire de fenêtres ou du traitement de texte lui-même, se sert alors de ce chiffre comme d'un index pour accéder à l'image (œil) correspondant au « T » qu'il dessine à l'écran. Le processus continue au fur et à mesure que l'utilisateur saisit de nouveaux caractères.

Le standard Unicode n'aborde directement que la question relative au codage et à la sémantique des caractères formant le texte et non les autres traitements effectués sur le texte. Dans le scénario précédent, le traitement de texte effectue peut-être une vérification orthographique sur le texte saisi et signale les fautes détectées. Le texteur pourrait encore insérer des passages à la ligne tous les *n* caractères. Un des principes fondamentaux d'Unicode stipule que le standard ne précise pas comment ces autres opérations doivent être réalisées pour peu que le codage et le décodage soient effectués correctement et que la sémantique des caractères soit préservée.

## Interprétation des caractères

Pour bien comprendre le rôle du standard Unicode dans le traitement textuel, il est indispensable de bien de bien distinguer l'identification d'une valeur de code de son affichage ou de son impression. Le caractère désigné par une valeur de code Unicode est une entité abstraite, par exemple : « LETTRE MAJUSCULE LATINE A » ou « CHIFFRE BENGALI CINQ ». L'image qui s'inscrit à l'écran, appelé glyphe ou œil, est une représentation visuelle du caractère.

Le standard Unicode ne définit pas les images de glyphes. En d'autres mots, le standard précise l'interprétation des caractères mais non le rendu de ceux-ci. En définitive, il revient au logiciel ou au matériel de rendu de l'ordinateur de déterminer l'apparence des caractères à l'écran. Le standard Unicode ne précise ni la taille, ni le format ni encore l'orientation des caractères à l'écran.

## Éléments textuels

Pour coder, traiter ou interpréter correctement un texte il faut disposer de définitions appropriées d'éléments textuels utiles et de règles de base d'interprétation de ce texte. La définition des éléments textuels dépend souvent du processus de traitement textuel. Ainsi, lors de la recherche d'un mot particulier ou d'un caractère écrit en écriture latine, on désire souvent ignorer la casse des caractères. Par contre, la vérification orthographique correcte d'un document nécessite de faire la distinction entre les minuscules et les majuscules.

Le standard Unicode ne spécifie pas ce qui constitue ou non un élément textuel pour les différents processus. Au contraire, il établit des éléments de codage, appelés valeurs de code. Une valeur de code, appelée communément un caractère, est un élément utile et fondamental pour tout traitement informatique textuel. Dans la plupart des cas, les valeurs de code correspondent aux éléments de texte les plus courants.

## 1.4 La norme ISO 10646 et le standard Unicode

Le standard Unicode est totalement compatible avec la norme internationale ISO/CEI 10646-1 : 2000, *Technologies de l'information — Jeu universel de caractères codés sur plusieurs octets (JUC) — Partie 1 : Architecture et plan multilingue de base*. Pendant l'année 1991, le consortium Unicode et l'Organisation internationale de normalisation (ISO) jugèrent souhaitable l'existence d'un seul code de caractères universel. À la suite de négociations officielles sur une convergence des deux codes, les deux répertoires fusionnèrent en janvier 1992. Depuis lors, une étroite collaboration et une liaison officielle entre les deux comités a permis d'assurer la synchronisation des amendements et additions aux deux normes de sorte que leurs répertoires respectifs et leurs codages sont aujourd'hui identiques.

La version 3.0 du standard Unicode est donc rigoureusement identique à l'ISO/CEI 10646-1 : 2000. Cette identité point par point s'applique à tous les caractères codés des deux normes y compris les caractères idéographique extrême-orientaux (également appelés caractères han). La norme ISO/CEI 10646 attribue à chaque caractère un nom et une valeur de code, le standard Unicode assigne les mêmes noms[5] et valeurs de code mais il fournit également d'importants algorithmes de mise en œuvre, des propriétés de caractères et d'autres renseignements sur la sémantique de ceux-ci.

Pour plus de détails sur le standard Unicode et l'ISO/CEI 10646, veuillez vous référer à l'annexe *C, Comparaison entre ISO/CEI 10646 et Unicode*.

## 1.5 Le consortium Unicode et le JTC1/SC2/GT2

Le consortium Unicode s'est constitué en personne morale en janvier 1991, sous le nom de Unicode inc., afin de faire connaître le standard Unicode en tant que système de codage international d'échange d'informations, de faciliter sa mise en œuvre et de s'assurer de la qualité des révisions ultérieures.

Pour atteindre ces buts, le consortium Unicode collabore avec l'Organisation internationale de normalisation (ISO/CEI/JTC1). Le consortium est membre de liaison de classe C auprès du JTC1/SC2 de l'ISO/CEI, il participe à la fois aux travaux du JTC1/SC2/GT2 (le groupe de travail du sous-comité au sein du JTC1 chargé du codage des jeux de caractères) et à ceux du GRI (groupe à rapporteur sur les idéophonogrammes) du GT2. Le consortium Unicode est une entreprise membre du *National Committee for Information Technology Standards*, comité technique L2 (NCITS/L2), un organisme de normalisation américain accrédité. En outre, les sociétés membres à part entière du consortium Unicode ont des représentants dans de nombreux pays qui collaborent avec les organismes de normalisation de ces pays.

Un certain nombre d'organismes sont membres de liaison auprès du consortium Unicode : le Centre de recherche et de développement informatique (CCID, Chine populaire), l'Internet engineering task force (IETF), la bibliothèque nationale Kongju (Chung-nam, Corée), le comité technique informatique (TCVN/TC1, Viêt-Nam) et le groupe de travail i18n du consortium World Wide Web (W3C).

Tout organisme ou particulier peut devenir membre du consortium quel que soit son pays d'origine pour peu que celui-ci soit en faveur du standard Unicode et qu'il soit disposé à

[5] Le standard Unicode ne précise pas de noms français, nous avons donc utilisé les noms officiels français de l'ISO/CEI 10646. Rappelons que les noms anglais n'ont aucune valeur normative pour l'ISO/CEI, seuls les numéros de caractères en ont une.

favoriser sa propagation et sa mise en œuvre généralisée. On retrouve parmi les membres de plein droit et les membres associés un large éventail d'entreprises et d'organismes dans le domaine de l'informatique. Les cotisations des membres sont la seule source de financement du consortium.

## Le comité technique d'Unicode

Le comité technique d'Unicode (UTC) est le groupe de travail au sein du consortium chargé de la création, de la mise à jour et de la qualité du standard Unicode. L'UTC contrôle toutes les données techniques fournies au consortium et prend des décisions quant au contenu relatif à ces données. Les membres de plein droit du consortium se prononcent sur les décisions de l'UTC. Les membres associés, les membres experts ainsi que les dirigeants du consortium font partie de l'UTC mais ne bénéficient pas de droit de vote. À l'invitation du président de l'UTC, d'autres personnes dans l'assistance peuvent participer aux discussions car le but de l'UTC est d'être un forum où l'on peut débattre librement de sujets techniques.