Encodage des caractères

Traitement Numérique Multilingue
Institut National des Langues Orientales - INaLCO
Paris, 75007 - FRANCE

Contenu

Le développement de l'Internet, la question du codage des caractères et les problèmes qui en ressortent font apparaitre tout l'enjeu d'une nécessité d'une harmonisation des encodages et d'une stantardisation des normes de codage. Parallèlement à ce processus de standardisation est apparu également un mouvement de normalisation industrielle. Parfois au prix d'un conflit de normes.

1 Définition

L'histoire d'Internet, et encore davantage son développement international, sont intimement liés à l'accessibilité aux ressources textuelles pour le plus grand nombre. Développé dans le monde anglo-saxon, l'Internet s'est d'abord exprimé en anglais, et c'est pour produire du texte dans cette langue que la première forme d'encodage de caractères a été créée.

Mais avant même d'aborder la notion d'encodage de caractères, il importe de se pencher sur la notion même de caractère. Selon la définition qu'en donne le Grand Dictionnaire Terminologique, un caractère est un *"elément de base d'un ensemble, employé conventionnellement pour exprimer une donnée de façon compréhensible par un être humain, ainsi que la forme codée [...] qui peut être traitée par un ordinateur"*.

Sommairement, un caractère est un glyphe, un des symboles d'un alphabet d'une langue naturelle, associé à une position binaire dans la mémoire centrale.

1.1 Jeu de caractères codés

Un jeu de caractères codés est une application corrélant les éléments d'un répertoire de caractères et un ensemble d'entiers positifs : à chaque élément du répertoire est assigné un code numérique unique, sa position de codage.

L'ensemble des positions de codage définit un espace de codage. Un caractère associé à une position de codage est dit caractère codé. On présente généralement les jeux de caractères codés sous la forme de tables (une ou plusieurs) que l'on appelle tables de caractères.

Une table de codage de caractères se présente sous la forme d'une liste de couples : chaque élément de la table désigne l'association d'une donnée numérique et d'un symbole permettant de coder une langue naturelle.

1.2 Formes d'encodage

Une forme d'encodage de caractères est une méthode (un algorithme) permettant de représenter les caractères d'un jeu de caractères codés en transformant leur code numérique en une séquence d'octets.

Dans le cas le plus simple, chaque caractère, par référence à une table de caractères, est mis en relation avec un entier compris entre 0 et 255 et cet en-

tier est utilisé tel quel en représentation binaire sur un format d'un octet. Cela n'est possible que dans le cas d'un répertoire restreint, comportant au maximum 256 éléments, ce qui nous amène à présenter le répertoire ASCII.

2 Le répertoire ASCII

Inventée en 1961 par l'américain Bob Bemer, la norme ASCII (American Standard Code for Information Interchange) est la norme de codage de caractères la plus connue dans le monde de l'Internet, et auss très certainement la plus compatible, puisqu'on la retrouve dans le noyau de toutes les normes ultérieures. C'est également la variante américaine du codage de caractères ISO/CEI 646. On a vu que le développement du net s'est produit dans un monde anglo-saxon, américain en particulier, et c'est l'ASCII qui a permis de rédiger les documents numériques dans cette langue, contenant les caractères nécessaires pour écrire en anglais. L'ASCII permet un codage sur 1 octet (7 bits), ce qui offre 128 positions binaires.

	0	1	2	3	4	5	6	7
0	NUL	DLE	space	0	@	P	`	p
1	SOH	DC1 XON	!	1	A	Q	a	q
2	STX	DC2	"	2	B	R	b	r
3	ETX	DC3 XOFF	#	3	C	S	c	s
4	EOT	DC4	$	4	D	T	d	t
5	ENQ	NAK	%	5	E	U	e	u
6	ACK	SYN	&	6	F	V	f	v
7	BEL	ETB	'	7	G	W	g	w
8	BS	CAN	(	8	H	X	h	x
9	HT	EM	)	9	I	Y	i	y
A	LF	SUB	*	:	J	Z	j	z
B	VT	ESC	+	;	K	[	k	{
C	FF	FS	,	<	L	\	l	|
D	CR	GS	-	=	M	]	m	}
E	SO	RS	.	>	N	^	n	~
F	SI	US	/	?	O	_	o	del

L'ASCII définit ainsi 128 caractères numérotés de 0 à 127 et codés en binaire de 0000000 à 1111111. Sept bits suffisent donc pour représenter un caractère codé en ASCII. Toutefois, les ordinateurs travaillant presque tous sur huit bits (un octet) depuis les années 1970, chaque caractère d'un texte en ASCII est stocké dans un octet dont le 8e bit est 0.

Les caractères de numéro 0 à 31 et le 127 ne sont pas affichables ; ils correspondent à des commandes de contrôle de terminal informatique. Le caractère numéro 32 est l'espace. Les autres caractères sont les chiffres arabes, les lettres latines majuscules et minuscules et quelques symboles de ponctuation.

Dans de telles conditions, il n'est pas possible de coder tous les signes diacritiques utilisés dans certaines langues (et notamment le français. Par ex. "é", "ç" ou "à"), mais uniquement les 26 lettres de l'alphabet latin standard, suffisant pour écrire de l'anglais.

2.1 L'ASCII élargi

Afin de répondre aux limitations de l'ASCII, en ayant le souci de l'internationalisation, on a procédé à une extension de l'ASCII, en proposant une table de caractères codée sur 8 bits. Le codage sur 8 bits permet de disposer de 256 caractères au lieu de 128 (28 au lieu de 27).

Les 128 premiers caractères correspondent à la table ASCII, tandis que les 128 suivants dépendent de la table d'encodage utilisée. Le jeu de caractères ISO-8859-1 (table de caractères codés sur 8 bits permettant l'écriture de nos caractères accentués français) est le jeu de caractère comportant nos caractères nationaux utilisé dans la hiérarchie française.

3 Une affaire de normes

Plusieurs normes sont apparues à la suite de l'ASCII. Toutefois, quelles qu'elles soient, toutes ces tables ont en commun de conserver dans leurs 128 premières positions le noyau de la norme ASCII (cf. Table ci-contre).

Parmi ces normes, il faut distinguer plusieurs normes selon qu'il s'agit de normes internationales ou de normes industrielles.

3.1. Normes internationales

Devant le multiplication des normes et la nécessité de les harmoniser, l'organisme de standardisation internationale (ISO) a décidé de procéder à

une standardisation des conventions de codage en établissant différentes normes ISO.

Nom de la table	Format de codage	Forme d'encodage
Iso 646 IRV	7 bits	Iso 646 (1 octet)
Iso8859-n pour n = [1;16] donc 16 tables	8 bits	Iso 8859 (1 octet)
Iso 10646 (UCS)	32 bits	UCS-4 (4 octets) UCS-2 (2 octets) UTF-16(2 octets ou 2 x 2 octets) UTF-8 (1 à 6 octets) UTF-7 (1 à 4 octets)

3.1.1 ISO 646

ISO 646 apparait comme la version standardisée de l'ASCII, la variante américiane de cette norme. Comme l'ASCII, ISO 646 est une norme de codage basé sur 7 bits. Jusqu'en 1991, ISO 646 et ASCII différait uniquement sur une position, et depuis cette date, les deux normes sont à l'identique.

3.1.2 ISO-8859

L'anglais étant l'une des seules langues utilisables avec l'ASCII/ISO 646, de nombreux organismes ont tenté de mettre sur pied des normes plus riches, et notamment l'établissement de la norme ISO 8859, qui est une extension sur 8 bits de l'ASCII. Passer de 7 à 8 bits permettait de doubler le nombre de caractères, passant à 256 caractères (les 128 premières positions étant réservés aux caractères codés par ASCII). Comme en Europe, il y a plus de 256 caractères différents utilisés, il a été décidé de décliner la norme ISO-8859 en 16 tables (ISO 8859-1/16).

Seul le codage ISO/IEC8859-1 (appelé aussi "LATIN-1") a été implanté généralement et est devenu un remplacement de facto de la norme ASCII en Europe. La norme ISO-latin1 permet d'écrire ces différentes langues: allemand, anglais, danois, espagnol, féroïen, finnois, français, islandais, italien, néerlandais, norvégien, portugais, suédois. La norme ISO-latin 2 est utilisée quant à elle pour le codage des langues de l'Europe orientale.

Cependant, même avec latin-1, le français n'avait toujours pas ses ligatures « œ » et « Œ » parce que les francophones n'avaient pas eu suffisamment besoin d'elles pour les exiger sur leurs claviers ; ni le Ÿ, bien que ce caractère soit utilisé dans des noms propres de famille et de lieux. Il faudra attendre l'ISO 8859-15 pour voir apparaitre ces caractères, en même temps que l'introduction du nouveau caractère de l'Euro (€).

3.2 Normes industrielles

Nom de la table	Format de codage	Forme d'encodage
Ebcdic (IBM)	8 bits	1 octet
Pages de codes de DOS 437, 850... (Microsoft)	8 bits	1 octet
Page de codes Windows 1250, 1251, 1252 (Microsoft) [Windows 1252 dite ANSI]	8 bits	1 octet
UNICODE (Consortium Unicode) Versions 1.*x* à 3.*x* Version 4.*x* à 5	 16 bits 20 bits	UCS-2 (2 octets) UTF-16(2 octets ou 2 x 2 octets) UTF-8 (1 à 6 octets) UTF-7 (1 à 4 octets

Parallèlement à cette volonté de standardisation apparait également de la part de certains géants de l'informatique (IBM, Microsoft) d'imposer leurs propres normes en matière d'encodage de caractères.

Toutefois, même une norme industrielle comme Windows 1252, de Microsoft, est en fait une extension de l'ISO/CEI 8859-1, différant malgré tout de l'ISO-8859-1 par l'utilisation de caractères imprimables, plutôt que des caractères de contrôle, dans la plage 80-9F, avec comme conséquences quelques problèmes de compatibilité. L'influence de certains constructeurs est telle que parfois, les normes industrielles font presque autorité face aux normes internationales.

Il faut aussi remarquer une norme qui se démarque de toutes les normes industrielles, il s'agit de Unicode. Le Consortium Unicode (qui réunit les plus grands noms de l'informatique, Adobe, Apple, IBM, Microsoft, Sun et Xerox) est une organisation à but non-lucratif qui coordonne le développement du projet. Elle a pour objectif de remplacer à terme les codages de caractères existants.

4 Unicode, ISO-10646 et BMP

4.1 UNICODE

"Le but d'UNICODE est de pouvoir fournir un codage non-ambigu sur 16 bits (jusqu'à la version 3.2, sur 20 bits depuis la version 4), qui n'a pas

besoin de séquences de contrôle, ni de méthode de compactage."[*]

Il doit permettre l'échange, le traitement et le visualisation des caractères du monde entier, partant du principe que l'ensemble des langues vivantes et mortes de l'humanité peuvent être codées de façon non-ambiguë en utilisant 2 octets (soit 16 bits).

Unicode est en quelque sorte une généralisation à double largeur d'ASCII. Il permet l'échange, le traitement et la visualisation des caractères utilisés par la plupart des langues vivantes: scripts latin, grec, cyrillic, arménien, hébreu, arabe, devanagari, bengali, gurmukhi, gujarati, oriya, tamul, télugu, kannada, malaysien, siamois, lao, géorgien, tibétain, kana, hangul, CJK (ensemble unifié des caractères idéographiques chinois, japonais, coréens).

Unicode définit un caractère (élément de codage d'un texte) en terme de 1 code + un nom mais ne définit aucun glyphe, c'est le dispositif qui utilise la table qui doit prendre en charge l'apparence du caractère. Actuellement, la table comprend environ 96 447 caractères (associations code-nom). Les caractères sont regroupés en «scripts» dans des blocs de codes.

Un script est un système de caractères ayant des propriétés communes. S'il y a un ordre habituel sur ces caractères, p.ex. ordre alphabétique, Unicode ordonne les caractère de telle sorte que cet ordre soit maintenu. Le projet UNICODE ne se contente pas de référencer, d'organiser et de classer les différents symboles des écritures. Il cherche à rationaliser leur utilisation et à établir des règles concernant leur manipulation. Il donne des recommandations et définit

- les caractères combinés : symboles complexes formés à partir de plusieurs symboles. UNICODE recense ces combinaisons et autorise leur définition par concaténation des caractères élémentaires, voire comme caractère unique à des fins de compatibilité avec les standard antérieur (c'est le cas des lettres diacritées du français).
- la normalisation des caractères afin d'établir des correspondances entre caractères de code points différents mais ayant la même interprétation ou la même fonction, entre caractères de casses (minuscule, majuscule et tittle-case) différentes pour rationaliser les conversions (p.ex. latin → cyrillic) et faciliter les comparaison et les tris.
- l'encodage des caractères

le Standard UNICODE est conçu pour être:

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

• **Efficace**. Le texte brut doit être facilement analysable : les logiciels ne doivent pas rechercher des séquences d'échappement ou maintenir une variable d'état, 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 un seul et même caractère (il n'y pas plus recours aux caractères d'échappement).

La version 3.1 du standard Unicode comprend 94.140 caractères issus de tous les systèmes d'écriture du monde. Cet ensemble est largement suffisant pour répondre aux besoins de communication moderne, permettant 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 71.394 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.

4.2 ISO 10646 et BMP

[*] Jacques ANDRE, Michel GOOSENS, Codage des caractères et multilinguisme : de l'ASCII à Unicode et ISO/CEI 10646, In : Cahiers Gutenberg, n° 20, juin 1995.

Le standard international ISO 10646 définit le jeu de caractères international, Universal Character Set (UCS). C'est un super-ensemble de tous les autres jeux de caractères standard. Il présente l'avantage de garantir une compatibilité réversible avec tous les autres jeux: il n'y aura donc aucune perte d'information si un texte est converti en UCS puis reconverti dans son code d'origine. UCS définit un jeu de caractères codés sur 31 bits.

À l'heure actuelle, les données Unicode peuvent être codées sous deux formes : une forme implicite de 16 bits (UTF-16) et une forme de 8 bits dénommée UTF-8 conçue pour faciliter son utilisation sur les systèmes ASCII préexistants.

UTF-8 est un codage où le caractère, au lieu d'être encodé sur 2 octets, l'est de manière variable sur 1 à 4 octetes, permettant une sérialisation plus efficace du texte. C'est le codage de base de XML, et donc le plus courant aujourd'hui.

UTF-16 permet de coder plus de 1 million de caractères, suffisant à coder tous les les caractères connus, y compris ceux utilisés par toutes les écritures historiques du globe.

Alors que UNICODE est toujours codé sur 2 octets, ISO 10646 code un ensemble de caractères sur plusieurs octets, soit quatre (UCS-4), soit deux (UCS-2). Dans le cas d'un codage sur 2 octets, on se réfère en réalité au codage du BMP.

Le sous-ensemble sur 16 bits de UCS s'appelle le BMP (Basic Multilingual Plan). La norme le définissant a été publiée en 1993 sous le nom de ISO 10646-1. UCS assigne à chaque caractère un code et un nom. Le code est un nombre en représentation hexadécimale. On a l'habitude lorsque l'on donne un code UCS (et Unicode) de le faire précéder de la lettre. Le nom est un nom standardisé (ex. U+0041 Latin capital letter A).

La structure du BMP est la suivante, il est réparti en 4 zones:

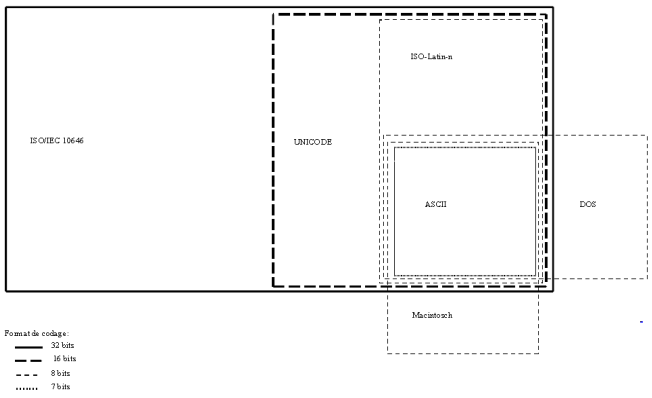
- zone A: alphabets arabe, arménien, cyrillique, grec, hangul, hébreu, indiens, kana, tha , ..., symboles diacritiques, symboles divers, éléments graphiques...

Les 256 premiers caractères correspondent aux caractères définis par ISO 8859-1 (ISO LATIN 1).

- zone I: idéogrammes (caractères chinois unifiés) Elle comporte environ 21000 caractères chinois unifiés de Chine, Corée et Japon. Ils ont été choisis dans les jeux de caractères définis par les normes GB2312 pour la Chine, Big-5 pour Taïwan, Jis X 0208 et JisX 0212 pour le Japon.
- zone O: ouverte (réservée pour extension, mais une partie est utilisée pour les hangul sous forme complète)
- zone R: réservée (pour usage privée et pour permettre les conversions de code).

UNICODE et ISO-10646 se développent conjointement. C'est au cours de l'année 1991 que le Consortium UNICODE et l'institution ISO émetterent l'idée d'un seul code de caractères universel. Moins d'un an plus tard, en janvier 1992, les deux répertoires fusionnèrent.

A partir de la version 3.0 d'UNICODE, les deux tables sont identiques. On peut donc affirmer une compatibilité totale du standard Unicode avec la norme internationale ISO-10646. Cette correspondance concerne tous les caractères codés des deux normes, comprenant également les idéogrammes utilisés en Extrême-Orient.



5 Notion d'interopérabilité et d'échange

Conçu sur le modèle du jeu de caractères ASCII, mais conscient de la nécessité de "penser global", UNICODE en surpasse les limitations rudimentaires, en codant tous les caractères utilisés par toutes les langues écrites du monde, soit plus d'un million de caractères réservés à cet effet. L'importance et la croissance des flux d'informations à

l'échelle planétaire rend impérieux le besoin de rendre interopérables et intercompréhensibles des logiciels qui doivent être utilisables n'importe où, quel que soit l'environnement linguistique ou culturel de l'utilisateur.

Unicode, et son pendant standardisé ISO-10646, est un mécanisme universel de codage de caractères qui définit une manière cohérente de coder des textes multilingues, facilitant l'échange de données textuelles et créant ainsi les prémisses pour tout logiciel international. Le standard UNICODE, en tant que codage implicite de HTML et XML, constitue un solide soubassement pour l'Internet. Obligatoire pour la plupart des nouveaux protocoles de l'Internet, mis en oeuvre dans tous les systèmes d'exploitation et langages informatiques modernes, Unicode est la base de tout logiciel qui doit fonctionner n'importe où dans le monde.

Au moyen d'Unicode, l'industrie informatique se trouve un moyen d'assurer la stabilité de ses données qui, en évitant la prolifération incontrôlée des jeux de caractères, augmente une interopérabilité et un échange de données au niveau mondial.

Toutefois, même si les progrès apportés par UNICODE sont immenses sur le chemin de l'interopérabilité, puisqu'aujourd'hui Unicode se trouve au cœur de tous les systèmes d'exploitation modernes (Windows, Mac OS, UNIX, etc.), Unicode n'est pas forcément encore reconnu et utilisé par toutes les applications, démontrant la difficulté sous-jacente à toute volonté de standardisation.

Bibliographie

Jacques ANDRE et Michel GOOSSENS. 1995. *Codage des caractères et multi-linguisme : de l'ASCII à UNICODE et ISO/IEC-10646*, Cahiers GUTenberg no20 — mai 1995

Patrick ANDRIES. *Introduction à Unicode et à l'ISO 10646*
[http://www.cairn.info/revue-document-numerique-2002-3-page-51.htm]

Steve FRECINAUX. *Introduction aux jeux de caractères*.
[http://openweb.eu.org/articles/jeux_caracteres/]

Grand Dictionnaire Terminologique
[http://www.oqlf.gouv.qc.ca/ressources/gdt.html]

Marie-Anne MOREAUX, *notes de cours (Principes de fonctionnement des ordinateurs)*. 2007-2008.