Codage des caractères

Jeux et codage de caractères

Nous différencions les jeux et les codages de caractères.

- Un **jeu de caractères** est une quantité déterminée de caractères (lettres, chiffres, symboles, etc.). Chaque caractère (en tant qu'unité abstraite, p. ex. «la majuscule latine A») est identifié de manière univoque par l'une des valeurs de codage qui lui est imputée (code de caractère).
- Un **codage de caractères** est une illustration de code de caractère sous une forme pouvant être représentée dans une mémoire numérique. Pour ce faire, une séquence de bits est attribuée à chaque code de caractère (qui représente un caractère abstrait précis).

Traditionnellement, les jeux et les codages de caractères coïncidaient: un caractère était représenté directement par une séquence de bits. Cela s'applique notamment à:

- ASCII 7 bits
- la famille de normes ISO 8859 (ISO-8859-1, etc.), en codage 8 bits.

Avec Unicode, les jeux et le codage de caractères sont séparés. Unicode définit en premier lieu les différents caractères des polices d'écriture respectives (jusqu'ici plus de 100 000) et leur attribue un dénommé *code point*, ou code caractère. Ce code caractère peut ensuite être converti de différentes manières dans une séquence de bits:

- UTF-32 (UTF signifie Unicode Transformation Format) dote chaque caractère d'une suite de 4 octets de 8 bits chacun. De ce fait, tous les caractères occupent la même place en mémoire et tous les caractères existants, ainsi que tous ceux qui doivent encore être définis peuvent être installés; cette représentation nécessite une grande capacité de mémoire.
- UCS-2 (2-byte Universal Character Set) dote chaque caractère de 2 octets de 8 bits chacun. De ce fait, tous les caractères occupent la même place en mémoire; 65 536 caractères peuvent être présentés, c'est-à-dire seulement le *Basic Multilingual Plane (BMP)* d'Unicode.
- UTF-16 résout ce problème en représentant les caractères du *BMP* dans un mot à 16 bits; les autres caractères Unicode dans deux mots à 16 bits.
- UTF-8 est un format de codage de caractères permettant l'optimisation de la place en mémoire. Ainsi, les 128 caractères ASCII sont représentés par un octet, ce qui permet d'assurer la compatibilité. Jusqu'à 4 octets seront utilisés pour les autres caractères du jeu. Il est recommandé d'utiliser UTF-8 pour optimiser la place en mémoire et assurer la compatibilité descendante avec ASCII.

Références

ASCII
American National Standards Institute (ANSI) X3.4-1967 (ASCII-1967)
ISO/IEC 646:1991, Technologie de l'information — Jeu ISO de caractères codés à 7 éléments pour l'échange d'information
https://www.iso.org/standard/4777.html
[payant]

ISO 8859
ISO/IEC 8859-1:1998, Technologie de l'information — Jeux de caractères graphiques codés sur un seul octet — Partie 1: Alphabet latin no. 1
https://www.iso.org/standard/28245.html
[payant]
http://std.dkuug.dk/jtc1/sc2/wg3/docs/n411.pdf
[gratuit, version bêta anglophone]

Unicode
Unicode 10.0.0
http://www.unicode.org/versions/Unicode10.0.0/
UTF-8
http://tools.ietf.org/html/rfc3629

Bibliographie

Spolsky, Joel: The Absolute Minimum Every Software Developer Absolutely, Positively Must Know About Unicode and Character Sets (No Excuses!)
http://www.joelonsoftware.com/articles/Unicode.html
Wikipédia, ISO 8859-1
http://fr.wikipedia.org/wiki/ISO_8859-1
Tero, Paul: Unicode, UTF8 & Character Sets: The Ultimate Guide
Smashing Magazine, 2012
https://www.smashingmagazine.com/2012/06/all-about-unicode-utf8-character-sets/

Unicode

UTF-8
http://www.utf-8.com/
Wikipedia: Comparatif de l'UTF-8 avec d'autres codages de caractères UNICODE
http://en.wikipedia.org/wiki/UTF-8#Advantages_and_disadvantages