



# 文字コード

私は宣言する。もしあなたが 21 世紀において仕事をしているプログラマであり、キャラクタ、キャラクタセット、エンコーディング、Unicode の基本について知らないのであれば、私はあなたをひっ捕まえて、潜水艦で 6 か月のたまねぎ剥きの刑に処する。—Joel Spolsky[44]

## B.1 文字コードとは

コンピュータが扱えるデータは、突き詰めればバイト列（ビット列）だけです。人間が何気なく扱っている「文字（キャラクタ）」も、コンピュータで扱うためにはバイト列で表さなければなりません。文字をバイト列で表す方法はいろいろ考えられますが、最も一般的なのは文字コードを利用する方法です。

文字コードとは、文字集合とその符号化方式の組のことです。符号化文字集合とも呼ばれます。文字集合は文字通り文字の集合<sup>1)</sup>、符号化方式は文字集合内の文字への番号の振

表 B.1 主要な文字コード

文字コード	文字集合	符号化方式	補足
ASCII			最も普及している文字コード。7 ビットの空間で 128 文字を表現する。制御文字と空白を除くと 94 文字。(図 6.7 p. 51)
ISO-8859-1			ASCII (94 文字) と西ヨーロッパ文字 (96 文字) を合わせたもの。Latin 1 とも呼ばれる。
JIS X 0201			ASCII (一部変更) にいわゆる「半角カナ」を追加したもの。特殊文字を除くと 158 文字 (空白を含む)。
ISO-2022-JP	JIS X 0201+ 208	ISO-2022	JIS X 0201 の半角カナは含まない。日本語のメールではこの文字コードがよく用いられる。ISO の規格ではない。JIS コードとも呼ばれるが、JIS 規格でもない。
EUC-JP	JIS X 0201+ 0208+ 0212	ISO-2022	Unix などによく使われていた文字コード。
Shift_JIS	JIS X 0201+ 0208	Shift_JIS	名称に JIS とあるが、JIS 規格ではない。
Windows-31J	JIS X 0201+ 0208+ 特殊文字	Shift_JIS	日本の PC のデファクト。間違っても Shift_JIS と呼ばれることが多い。
UTF-8	UCS	UTF-8	Unicode の全ての文字を扱えるが、漢字は 1 文字あたり 3 バイト以上必要。
UTF-16	UCS	UTF-16	2 バイトあるいは 4 バイトで 1 文字を表す。

1) 集合とは言っても番号を振って定義されているものが多いです。そういう意味では文字コードだとも言えます。このあたりは厳密に使い分けられているわけではありません。たとえば、ASCII は文字コードの意味でも文字集合の意味でも使われます。