符号化文字集合とは、要するに文字コードのことです。文字とコードとの対応付けの定義のことをいいます。この文脈における「コード」とは、ディジタルコンピュータを前提とすると一般的に、ビットを組み合わせたバイト (列) を意味します。

規格における定義

JIS では下記のような定義がなされています。(JIS X 0213:2000 より)

> 符号化文字集合 (coded character set), 符号 (code)　文字集合を定め , かつその集合内の文字とビット組合せとを 1 対 1 に関係付ける , あいまいでない規則の集合。

これは ISO/IEC 646 や ISO/IEC 8859 といった国際規格における定義と同等の内容です。例えば ISO/IEC 646:1991 では下記のように定義されています。ASCII を定める ANSI X3.4 でも同じです。

> coded character set; code : A set of unambiguous rules that establishes a character set and the one-to-one relationship between the characters of the set and their bit combinations.

「符号 (code)」という言葉が「符号化文字集合」と同じとして定義されていることから、一般的にいう「文字コード」も「符号化文字集合」と同じとみなして用いることができます。

ただ、「コード」という言葉は、ここで定義されているような体系のことを指すこともあれば、コードによって定義される個々の値のことを指すこともあります。それが紛らわしいときは、前者を「コード系」、後者を「コード値」のように言って区別することができます。

JIS X 0213 が定義する符号化文字集合

JIS X 0213 では、漢字集合 1 面と漢字集合 2 面という 2 つの符号化文字集合を定義しています。

2 つ以上の符号化文字集合を組み合わせたものもまた、上記の定義における符号化文字集合です。JIS X 0213 は「符号化文字集合」の節において、6 種類の符号化文字集合を定義しています。6 種類の内訳は、

- JIS X 0213 を単体で用いるか、または ISO/IEC 646 国際基準版 (ASCII と同等) あるいは JIS X 0201 の 1 バイトコードと組み合わせて用いるか
- 7 ビットか 8 ビットか

という 2 つの軸によって整理されます。(参考 :「JIS X 0213 のあまり代表的でないかもしれない符号化方式」)

また、附属書 1 から 3 において、Shift_JIS-2004, ISO-2022-JP-2004, EUC-JIS-2004 を定義していますが、これらも上記の定義における符号化文字集合です。こうしたものは、IETF などの非標準の文書では文字符号化方式とされることがしばしばありますが、用語の定義に違いによるものです。

符号化文字集合についての誤解

符号化文字集合を、符号化表現を与えられていない単なる文字の集合とした解説がたまにありますが、上記 JIS や ISO の定義からも分かるとおり、これは誤解です。

集合の各要素 (文字) に符号化表現が与えられているからこそ「符号化文字集合」と呼ばれるのであって、そうでなければただの「文字集合」です。符号化されていない文字集合としては、例えば、常用漢字表や、平仮名の集合 (「いろは」47 文字など)、ラテンアルファベット 26 文字の集合、などを挙げることができます。

また、文字コードを定める JIS や ISO 以外の業界団体の発行する文書には、符号化文字集合を文字と整数の対応付けと記したものがありますが、これも不適当です。JIS や ISO, ANSI (ASCII) といった世界の文字コード標準では上に引用したように文字とビット組合せの対応付けとして定義されています。ビット組合せは 2 進法の整数と解釈することができますが、その方法は一通りではありません (1 の補数、2 の補数など)。

参考

- ISO/IEC 646, ISO/IEC 8859-1, JIS X 0208, JIS X 0213 等の符号化文字集合規格の用語定義
- プログラマのための文字コード技術入門 第 1 章

関連項目

- ISO/IEC 2022 - 符号化文字集合の構造と拡張法を定める国際標準。