Unicode は、Unicode コンソーシアムという業界団体が策定している文字コードの国際規格です。世界中の文字を単一の文字コード (符号化文字集合) で扱うことをねらいとしています。

JIS X 0213 との関係

Unicode は、バージョン 3.2 で、JIS X 0213 の文字を全て取り込みました。

JIS X 0213 と Unicode のコード変換表は当サイトで入手できます。

- JIS X 0213 のコード対応表

Unicode で JIS X 0213 の文字を扱うには、以下のようにいくつか問題があります。

■結合文字の問題

Unicode では、JIS X 0213 の全ての文字に対し単一の符号位置が割り当てられているわけではありません。結合文字を用いて複数の符号位置の並びで表現できる文字については、単一の符号位置は与えられていません。

例えば、鼻濁音を表すのに使われる半濁点つきの「か」は、JIS X 0213 では一つの独立した符号位置が与えられています (面区点 1-04-87)。しかし Unicode では、「か」+「合成用半濁点 (U+309A)」という 2 つの符号位置の並びによって表現する必要があります。

■互換漢字の問題

JIS X 0213 で追加された漢字のいくつかは、Unicode では CJK 互換漢字として扱われています。これは、従来の CJK 統合漢字に包摂されている字体のものです。

例えば、「神」の示へんが「ネ」でなく「示」の字体は、Unicode では包摂されています (つまり区別せず同じ符号位置で表す)。JIS X 0213 はこれに独立した符号位置を与えましたが (面区点 1-89-28)、Unicode では CJK 互換漢字として、JIS X 0213 との往復変換用との扱いで追加されています (U+FA19)。これは、Unicode 正規化の処理を適用すると、対応する CJK 統合漢字に移されてしまいます。

■サロゲートペアの問題

JIS X 0213 の一部の漢字は BMP でなく面 02 に追加されています。このため、UTF-16 でサロゲートペアによって、4 バイトで 1 文字を表します。

また UTF-8 では漢字は通常 3 バイトですが、面 02 の漢字は 4 バイトの長さになります。UTF-8 を扱うソフトウェアが 4 バイトを正しく扱えるか注意が必要です。

参照情報

Unicode 仕様書は Unicode コンソーシアムのウェブサイトから PDF 形式で入手できます。同サイトでは各文字の属性情報などを機械可読形式で記したテキストファイルも配布されています。

関連項目

- ISO/IEC 10646 - ISO と IEC による、Unicode と同等の文字コードの公的標準。
- BMP
- UTF-8
- UTF-16