

## Unicode って何ですか？

Windows や UNIX、Java 上において多言語文字をあらわす文字コードです。

特徴として下記があります。

- ・ エンコーディング(符号化方式)に UTF-7, UTF-8, UTF-8N, UTF-16BE, UTF-16LE, UTF-32 と数種類存在する。
- ・ **基本的**には、文字あたり 2 バイトを使用する。2 バイトなので 65536 通り (0x0000~0xFFFF) のビットを表現できます。この約 6 万字で世界中の文字を表現しようというのが Unicode の本来の思想だった。
- ・ 「未登録言語、追加の漢字(合計数万)も追加してほしい」と要望があり、従来の 2 バイト(65536 文字)では不足してしまった。解決策として**サロゲートペア**という方法が導入された。「1 文字=2 バイト」の基本を維持し、一部の文字は「1 文字=4 バイト」にする方法が採用された。
- ・ 複数バージョンが存在し、互換性が無い場合もある。現在まで Unicode 1.0.0~5.2.0 が存在する。

## エンコーディング(符号化方式)

- ・ UTF-7           あまり使用しないので省略
- ・ UTF-8           可変長(1~4 バイト)で表現する文字コード  
ASCII に対して上位互換(ASCII 文字と互換性を持たせるために、ASCII と同じ部分は 1 バイト、その他の部分を 2~4 バイトで符号化)  
UTF-16/32 との変換・逆変換が容易  
インターネットではもっとも一般的に利用されている  
(日本国内でのみ) **BOM**※がついている

※ **BOM**(Byte Order Mark)

UTF-8 であることが識別できるようにデータの先頭に  
付与された 0xEF BB BF の 3 バイトのこと

-参考-

- ・ Windows のメモ帳で作成した「Unicode テキスト」には BOM が付与される。
- ・ Internet Explorer では、BOM のついていない UTF-8 の文書を読み込むと(日本語版の場合) S-JIS と誤認する

- ・ UTF-8N            **BOM**(Byte Order Mark)がついていない。
  
- ・ UTF-16LE        1文字が、16ビットの符号単位が1つまたは2つで符号化される。  
                     サロゲートペア[後述]により、1文字が、2バイトだったり4バイト  
                     だったりする。  
                     LEは、リトルエンディアンの略(下位バイトからメモリに並べるため、  
                     2バイト毎ひっくり返って並べられます。)  
                     BOMが付与される場合、0x"FFFE"となる。
  
- ・ UTF-16BE        BEは、ビッグエンディアンの略(上位バイトからメモリに並べる。)  
                     BOMが付与される場合、0x"FEFF"となる。
  
- ・ UTF-32           1文字が、32ビットの符号単位が1つで符号化される。  
                     Unicode 3.1より実装された。  
                     UTF-32は、テキストファイルで使用されることは少なく、主にシステム  
                     のメモリ上での管理や、データベース等で使用される。

## サロゲートペア

従来のUnicodeでは未使用のだった0x"D800"~"0xDBFF"(1024通り)を「上位サロゲート」、0x"DC00"~0x"DFFF"(1024通り)を「下位サロゲート」と規定し、「上位サロゲート+下位サロゲート」の4バイトで文字を表現する方法です。

「上位サロゲート」も「下位サロゲート」も従来のUnicodeでは未使用の領域なので、以前のUnicodeの文字コードと重複することはありません。

このサロゲートペアの導入により1024×1024=1048576字の領域が追加されることになりました。

Windows Vista(JIS2004)で追加された907字の中でサロゲートペアを使用する文字は304字あります。

