UTF-8 は、Unicode の符号化方式のひとつです。Unicode の符号位置 (大まかに言うと文字に対応) ひとつにつき、1 バイトから 4 バイトまでの長さを取る可変長のコードです。

JIS X 0213 との関係

現在の Unicode は JIS X 0213 の文字を全て含んでいるため、UTF-8 で JIS X 0213 の文字全てを符号化することが可能です。

ただし、いくつかの文字は BMP 外にあるため、UTF-8 では 4 バイトに対応していることが必要です。BMP の範囲は UTF-8 で 3 バイトまでで表現でき、なおかつ昔は BMP にしか文字が割り当てられていなかったため、UTF-8 といっても 3 バイトまでしか対応していないことがあります。例えばMySQL では utf8 と指定すると3バイトまでしか対応しません。4バイトに対応するには utf8mb4 と指定する必要があります。

コード変換

iconv コマンドで Shift_JIS-2004 のテキストデータを UTF-8 に変換するには次のようにします。SJIS から UTF-8 に変換するときは常にこの指定を用いるのがおすすめです。

```
iconv -f SHIFT_JISX0213 -t UTF-8 < sjis.txt > utf8.txt
```

反対に、UTF-8 から Shift_JIS-2004 にするには下記のようにします。

```
iconv -f UTF-8 -t SHIFT_JISX0213 < utf8.txt > sjis.txt
```

関連項目

- Unicode
- ISO/IEC 10646
- UTF-16