JAPLA 研究会資料 2011/3/26

Ｊによるシステム処理－その４
日本語文字コード（シフト JIS、Unicode、UTF-8）の中身を見る

西川　利男

１．文字コードとは

コンピュータ上で文字情報を表すには次のように行う。人間にも読める文字の並びはフォントと呼ばれる一種の図形を並べることになるが、コンピュータのデータとしてはその文字の所在を示すアドレスの値を用い、これが文字のコードである。

英字と数字のコードについては、１バイトで行われ、いわゆる ASCII コードである。日本語文字のコードは、ふつう２バイトが用いられるが、これまでいろいろな方式が行われて来た。すなわち、JIS、 Shift_JIS、EUC などがある。最近は多言語用文字コードである UTF-8、UTF-16 なども使用される。

このようないろいろな方式は、開発したコンピュータ・メーカ、さらには JIS、ISO など国内、海外との仕様の調整、変更などの反映であるが、ユーザを悩ましている。

Ｊによるシステム処理の適用の例として、いろいろな日本語文字コードの違いを、ファイルの中身を実際に見てみることによってしらべてみた。

メモ帳(NotePad)のファイルとして、以下のように簡単な英数文字、日本語文字の文書データを文字コードを変えて作り、それを比較してみる。

まず、

```
This is a test.
これは日本語のデータです。
1 2 3 4 5 6 7 8 9 10
```

２．ANSI コード (Shift_JIS)

Shift_JIS はこれまで最もよく使われてきた。

```
== ANSI code ======================================
   debug '¥Nihongo_Code¥nsentence.txt'
File Length = 67(=43 hex) bytes
Dump Address (hex): from 0000 to 004F
*** Dump go_on ? (y or q)
y
     -0 -1 -2 -3 -4 -5 -6 -7 -8 -9 -A -B -C -D -E -F
000: 54 68 69 73 20 69 73 20 61 20 74 65 73 74 2E 0D This is a test._

001: 0A 82 B1 82 EA 82 CD 93 FA 96 7B 8C EA 82 CC 83 _         {
        こ    れ    は    日    本    語    の    デ
002: 66 81 5B 83 5E 82 C5 82 B7 81 42 0D 0A 31 20 32 f [ ^     B__1 2
        ー    タ    で    す    。
003: 20 33 20 34 20 35 20 36 20 37 20 38 20 39 20 31  3 4 5 6 7 8 9 1
004: 30 0D 0A                                       0__
*** end ***
```

３．UTF-8 コード

UTF コードとは、日本語に限らず世界中のいろいろな言語の文字を表すため、最近多言語用のコードとして使われてきている。

英数文字は１バイトで ASCII と同じだが、最初に英数文字であることを示す３バイトを付加する。

日本語文字は１文字を３バイトで表す。最初の１バイトでは漢字、ひらがなを区別する。後の２バイトが狭義のコードである。

原則として、３バイトで表すという方式である。

```
== utf8 code ======================================
   debug '¥Nihongo_Code¥nsen_utf8.txt'
File Length = 83(=53 hex) bytes
Dump Address (hex): from 0000 to 005F
*** Dump go_on ? (y or q)
y
     -0 -1 -2 -3 -4 -5 -6 -7 -8 -9 -A -B -C -D -E -F
000: EF BB BF 54 68 69 73 20 69 73 20 61 20 74 65 73    This is a tes

001: 74 2E 0D 0A E3 81 93 E3 82 8C E3 81 AF E6 97 A5 t.__
                こ       れ       は       日
002: E6 9C AC E8 AA 9E E3 81 AE E3 83 87 E3 83 BC E3
     本       語       の       デ       ー       タ
003: 82 BF E3 81 A7 E3 81 99 E3 80 82 0D 0A 31 20 32           __1 2
           で       す       。
004: 20 33 20 34 20 35 20 36 20 37 20 38 20 39 20 31  3 4 5 6 7 8 9 1
005: 30 0D 0A                                        0__
*** end ***
```

４．ユニコード（原則的ユニコード）

ユニコードは英数文字(記号も含めて)、日本語文字の両方ともに２バイトで表すことを原則にしている。

```
== unicode big endian =============================
   debug '¥Nihongo_Code¥nsen_unibig.txt'
File Length = 110(=6E hex) bytes
Dump Address (hex): from 0000 to 006F
*** Dump go_on ? (y or q)
y
     -0 -1 -2 -3 -4 -5 -6 -7 -8 -9 -A -B -C -D -E -F
000: FE FF 00 54 00 68 00 69 00 73 00 20 00 69 00 73    T h i s   i s

001: 00 20 00 61 00 20 00 74 00 65 00 73 00 74 00 2E    a   t e s t .

002: 00 0D 00 0A 30 53 30 8C 30 6F 65 E5 67 2C 8A 9E  _ _0S0 0oe g,
                こ    れ    は    日    本    語
```

```
003: 30 6E 30 C7 30 FC 30 BF 30 67 30 59 30 02 00 0D 0n0 0 0 0g0Y0  _
     の    デ    ー    タ    で    す    。
004: 00 0A 00 31 00 20 00 32 00 20 00 33 00 20 00 34  _ 1   2   3   4

005: 00 20 00 35 00 20 00 36 00 20 00 37 00 20 00 38    5   6   7   8

006: 00 20 00 39 00 20 00 31 00 30 00 0D 00 0A          9   1 0 _ _
*** end ***
```

５．ユニコード(慣用のユニコード)

慣用されるユニコードは、コードの２バイト値は高バイトと低バイトで逆順になっている。

```
== unicode ======================================
   debug '¥Nihongo_Code¥nsen_unicode.txt'
File Length = 110(=6E hex) bytes
Dump Address (hex): from 0000 to 006F
*** Dump go_on ? (y or q)
y
     -0 -1 -2 -3 -4 -5 -6 -7 -8 -9 -A -B -C -D -E -F
000: FF FE 54 00 68 00 69 00 73 00 20 00 69 00 73 00   T h i s   i s

001: 20 00 61 00 20 00 74 00 65 00 73 00 74 00 2E 00   a   t e s t .

002: 0D 00 0A 00 53 30 8C 30 6F 30 E5 65 2C 67 9E 8A _ _ S0 0o0 e,g
                こ    れ    は    日    本    語
003: 6E 30 C7 30 FC 30 BF 30 67 30 59 30 02 30 0D 00 n0 0 0 0g0Y0 0_
     の    デ    ー    タ    で    す    。
004: 0A 00 31 00 20 00 32 00 20 00 33 00 20 00 34 00 _ 1   2   3   4

005: 20 00 35 00 20 00 36 00 20 00 37 00 20 00 38 00   5   6   7   8

006: 20 00 39 00 20 00 31 00 30 00 0D 00 0A 00         9   1 0 _ _
*** end ***
```

６．WORDファイル文書

WORDの初期設定としては、上の慣用ユニコードになっている。なお、WORDのファイルはメモ帳と比べると相当大きいが、前半の部分には、WORDのいろいろな設定などが書かれているらしい(?)。これは、後半の文字コードの部分を取り出したものである。

```
   (<'A00') debug '¥Nihongo_Code¥nsentence.doc'
¥Nihongo_Code¥nsentence.doc
File Length = 24576(=6000 hex) bytes
Dump Address (hex): from 0A00 to 5FFF
*** Dump go_on ? (y or q)
y
     -0 -1 -2 -3 -4 -5 -6 -7 -8 -9 -A -B -C -D -E -F
0A0: 54 00 68 00 69 00 73 00 20 00 69 00 73 00 20 00 T h i s   i s

0A1: 61 00 20 00 74 00 65 00 73 00 74 00 2E 00 0D 00 a   t e s t . _

0A2: 53 30 8C 30 6F 30 E5 65 2C 67 9E 8A 6E 30 C7 30 S0 0o0 e,g  n0 0
     こ    れ    は    日    本    語    の    デ
0A3: FC 30 BF 30 67 30 59 30 02 30 0D 00 31 00 20 00  0 0g0Y0 0_ 1
     ー    タ    で    す    。
0A4: 32 00 20 00 33 00 20 00 34 00 20 00 35 00 20 00 2   3   4   5

0A5: 36 00 20 00 37 00 20 00 38 00 20 00 39 00 20 00 6   7   8   9

0A6: 31 00 30 00 0D 00 0D 00 00 00 00 00 00 00 00 00 1 0 _ _

0A7: 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00
go_on ? (y or q)
q
*** quit ***
```