

# UnicodeをUTF-8やUTF-16に変換する方法

🔗 UTF-8, 文字コード, unicode, UTF-16

## Unicodeの符号

Unicodeは面（0～16）、句（0～255）、点（0～255）の符号空間を持っており、Unicode符号位置を文章中などに記す場合などは **U+** の後に16進数でその値を続けることで表します。

### Unicodeの符号空間

	面	句	点
10進数	0～16	0～255	0～255
16進数	0x00～0x10	0x00～0xFF	0x00～0xFF

例えば **a** は **U+0061**、**あ** は **U+3042**、**鰻** は **U+29E7D** と表記されます。

文字	面	句	点
a	0x00	0x00	0x61
あ	0x00	0x30	0x42
鰻	0x02	0x9E	0x7D

Unicodeの各面に収録されている文字は以下となります。

面	符号位置	名称	略称	収録されている主な文字
第0面	U+0000 - U+FFFF	Basic Multilingual Plane（基本多言語面）	BMP	基本的な文字。
第1面	U+10000 - U+1FFFF	Supplementary Multilingual Plane（追加多言語面）	SMP	古代文字や記号・絵文字類など。
第2面	U+20000 - U+2FFFF	Supplementary Ideographic Plane（追加漢字面）	SIP	漢字専用領域。
第3面	U+30000 - U+3FFFF	Tertiary Ideographic Plane（第三漢字面）	TIP	追加漢字面に入りきらなかった漢字。また、将来的には古代漢字や甲骨文字などが収録される予定。
第4面	U+40000 - U+4FFFF	未使用	-	-
第5面	U+50000 - U+5FFFF	未使用	-	-
第6面	U+60000 - U+6FFFF	未使用	-	-
第7面	U+70000 - U+7FFFF	未使用	-	-
第8面	U+80000 - U+8FFFF	未使用	-	-
第9面	U+90000 - U+9FFFF	未使用	-	-
第10面	U+A0000 - U+AFFFF	未使用	-	-
第11面	U+B0000 - U+BFFFF	未使用	-	-
第12面	U+C0000 - U+CFFFF	未使用	-	-
第13面	U+D0000 - U+DFFFF	未使用	-	-
第14面	U+E0000 - U+EFFFF	Supplementary Special-purpose Plane（追加特殊用途面）	SSP	制御コード専用領域。
第15面	U+F0000 - U+FFFFF	Private Use Plane（私用面）	PUP	BMPの U+E000 - U+F8FF の領域の拡張。
第16面	U+100000 - U+10FFFF	Private Use Plane（私用面）	PUP	BMPの U+E000 - U+F8FF の領域の拡張。

## UnicodeをUTF-16に変換

### 変換手順

- （1）文字のUnicode符号位置を  $n$  とします。
- （2） $n$  が 0x10000 より小さい（第0面）場合、 $n$  を16ビットの符号なし整数として表現して終了です。 $n$  が 0x10000 以上（サロゲートペア）の場合、（3）に進みます。

- （3） $n' = n - 0x10000$ 、 $w_1 = 0xD800$ （1101 1000 0000 0000）、 $w_2 = 0xDC00$ （1101 1100 0000 0000）とします。

こうすると、 $n'$  は20ビット以内で表現可能となり、 $w_1$ 、 $w_2$  は下位10ビットが0となります。

$$n' = \text{yyyy yyyy yyxx xxxx xxxx}$$

$$w_1 = 1101\ 1000\ 0000\ 0000$$

$$w_2 = 1101\ 1100\ 0000\ 0000$$

- （4） $n'$  の各ビットを  $w_1$  と  $w_2$  にそれぞれ10ビットずつ割り振ります。

$$n' = \text{yyyy yyyy yyxx xxxx xxxx}$$

$$w_1 = 1101\ 10\text{yy yyyy yyyy}$$

$$w_2 = 1101\ 11\text{xx xxxx xxxx}$$

$w_1$  が上位サロゲート、 $w_2$  が下位サロゲートとなります。

### UTF-16の変換例（あ（U+3042）の場合）

- （1） $n = 0x3042$
- （2） $n = 0x3042$  は 0x10000 より小さいため、「あ（U+3042）」のUTF-16のコードは「**0x3042**」となります。

### UTF-16の変換例（鰻（U+29E7D）の場合）

- （1） $n = 0x29E7D$
- （2） $n = 0x29E7D$  は 0x10000 以上。
- （3） $n' = n - 0x10000 = 0x29E7D - 0x10000 = 0x19E7D$ （0001 1001 1110 0111 1101）
- $$w_1 = 1101\ 1000\ 0000\ 0000$$
- $$w_2 = 1101\ 1100\ 0000\ 0000$$
- （4） $n' = n - 0x10000 = 0x29E7D - 0x10000 = 0x19E7D$ （0001 1001 1110 0111 1101）
- $$w_1 = 1101\ 1000\ 0110\ 0111 = \text{0xD867}$$
- $$w_2 = 1101\ 1110\ 0111\ 1101 = \text{0xDE7D}$$

UTF-16BEの場合、**0xD867 0xDE7D**

UTF-16LEの場合、**0xDE7D 0xD867**

## UnicodeをUTF-8に変換

### 変換手順

- （1）Unicodeの符号位置の範囲を調べます。

	符号位置（16進）	UTF-8のビット列（2進）
範囲1	U+0000 - U+007F	0xxx xxxx
範囲2	U+0080 - U+07FF	110x xxxx 10xx xxxx
範囲3	U+0800 - U+FFFF	1110 xxxx 10xx xxxx 10xx xxxx
範囲4	U+10000 - U+10FFFF	1111 0xxx 10xx xxxx 10xx xxxx 10xx xxxx

- （2）Unicodeの符号位置のビット列を求めます。

- （3）調べた範囲のUTF-8のビット列（2進）のxにUnicodeの符号のビット列を当てはめた値がUTF-8の値となります。

### UTF-8の変換例（a（U+0061）の場合）

- （1）a（U+0061）は「範囲1」となります。
- （2）U+0061のビット列は「0**110 0001**」となります。
- （3）「範囲1」のUTF-8のビット列（2進）「0xxx xxxx」に当てはめると、「0110 0001」となり、16進数にすると「**0x61**」となります。

### UTF-8の変換例（®（U+00AE）の場合）

- （1）®（U+00AE）は「範囲2」となります。
- （2）U+00AEのビット列は「0**000 1010 1110**」となります。
- （3）「範囲2」のUTF-8のビット列（2進）「110x xxxx 10xx xxxx」に当てはめると、「1100 0010 1010 1110」となり、16進数にすると「**0xC2AE**」となります。

### UTF-8の変換例（あ（U+3042）の場合）

- （1）あ（U+3042）は「範囲3」となります。
- （2）U+3042のビット列は「0**011 0000 0100 0010**」となります。
- （3）「範囲3」のUTF-8のビット列（2進）「1110 xxxx 10xx xxxx 10xx xxxx」に当てはめると、「1110 0011 1000 0001 1000 0010」となり、16進数にすると「**0xE38182**」となります。

### UTF-8の変換例（鰻（U+29E7D）の場合）

- （1）鰻（U+29E7D）は「範囲4」となります。
- （2）U+29E7Dのビット列は「0**0010 1001 1110 0111 1101**」となります。
- （3）「範囲4」のUTF-8のビット列（2進）「1111 0xxx 10xx xxxx 10xx xxxx 10xx xxxx」に当てはめると、「1111 0000 1010 1001 1011 1001 1011 1101」となり、16進数にすると「**0xF0A9B9BD**」となります。

## UnicodeをUTF-32に変換

Unicodeの符号位置の値を4バイト固定幅にしたものがそのままUTF-32となります。

### 参考

プログラマのための文字コード技術入門（ISBN978-4-7741-4164-0）