NAME

euc - EUC encoding of wide characters

SYNOPSIS

ENCODING "EUC"

VARIABLE *len1 mask1 len2 mask2 len3 mask3 len4 mask4 mask*

DESCRIPTION

EUC implements a system of 4 multibyte codesets. A multibyte character in the first codeset consists of *len1* bytes starting with a byte in the range of 0x00 to 0x7f. To allow use of ASCII, *len1* is always 1. A multibyte character in the second codeset consists of *len2* bytes starting with a byte in the range of 0x80-0xff excluding 0x8e and 0x8f. A multibyte character in the third codeset consists of *len3* bytes starting with the byte 0x8e. A multibyte character in the fourth codeset consists of *len4* bytes starting with the byte 0x8f.

The *wchar_t* encoding of **EUC** multibyte characters is dependent on the *len* and *mask* arguments. First, the bytes are moved into a *wchar_t* as follows:

byte0 << ((*len*N-1) * 8) | byte1 << ((*len*N-2) * 8) | ... | byte*len*N-1

The result is then ANDed with ~*mask* and ORed with *maskN*. Codesets 2 and 3 are special in that the leading byte (0x8e or 0x8f) is first removed and the *lenN* argument is reduced by 1.

For example, the ja_JP.eucJP locale has the following *VARIABLE* line:

```
VARIABLE        1 0x0000 2 0x8080 2 0x0080 3 0x8000 0x8080
```

Codeset 1 consists of the values 0x0000 - 0x007f.

Codeset 2 consists of the values who have the bits 0x8080 set.

Codeset 3 consists of the values 0x0080 - 0x00ff.

Codeset 4 consists of the values 0x8000 - 0xff7f excluding the values which have the 0x0080 bit set.

Notice that the global *mask* is set to 0x8080, this implies that from those 2 bits the codeset can be determined.

SEE ALSO

localedef(1), setlocale(3)