

NAME

multibyte - multibyte and wide character manipulation functions

LIBRARY

Standard C Library (libc, -lc)

SYNOPSIS

#include <limits.h>

#include <stdlib.h>

#include <wchar.h>

DESCRIPTION

The basic elements of some written natural languages, such as Chinese, cannot be represented uniquely with single C *chars*. The C standard supports two different ways of dealing with extended natural language encodings: wide characters and multibyte characters. Wide characters are an internal representation which allows each basic element to map to a single object of type *wchar_t*. Multibyte characters are used for input and output and code each basic element as a sequence of C *chars*. Individual basic elements may map into one or more (up to MB_LEN_MAX) bytes in a multibyte character.

The current locale (setlocale(3)) governs the interpretation of wide and multibyte characters. The locale category LC_CTYPE specifically controls this interpretation. The *wchar_t* type is wide enough to hold the largest value in the wide character representations for all locales.

Multibyte strings may contain ‘shift’ indicators to switch to and from particular modes within the given representation. If explicit bytes are used to signal shifting, these are not recognized as separate characters but are lumped with a neighboring character. There is always a distinguished ‘initial’ shift state. Some functions (e.g., mblen(3), mbtowc(3) and wctomb(3)) maintain static shift state internally, whereas others store it in an *mbstate_t* object passed by the caller. Shift states are undefined after a call to setlocale(3) with the LC_CTYPE or LC_ALL categories.

For convenience in processing, the wide character with value 0 (the null wide character) is recognized as the wide character string terminator, and the character with value 0 (the null byte) is recognized as the multibyte character string terminator. Null bytes are not permitted within multibyte characters.

The C library provides the following functions for dealing with multibyte characters:

Function	Description
mblen(3)	get number of bytes in a character
mbrlen(3)	get number of bytes in a character (restartable)

`mbrtowc(3)` convert a character to a wide-character code (restartable)

`mbsrtowcs(3)`

convert a character string to a wide-character string (restartable)

`mbstowcs(3)` convert a character string to a wide-character string

`mbtowc(3)` convert a character to a wide-character code

`wcrtomb(3)` convert a wide-character code to a character (restartable)

`wcstombs(3)` convert a wide-character string to a character string

`wcsrtombs(3)`

convert a wide-character string to a character string (restartable)

`wctomb(3)` convert a wide-character code to a character

SEE ALSO

`localedef(1)`, `setlocale(3)`, `stdio(3)`, `big5(5)`, `euc(5)`, `gb18030(5)`, `gb2312(5)`, `gbk(5)`, `mkanji(5)`, `utf8(5)`

STANDARDS

These functions conform to ISO/IEC 9899:1999 ("ISO C99").