Null-terminated multibyte strings

A null-terminated multibyte string (NTMBS), or "multibyte string", is a sequence of nonzero bytes followed by a byte with value zero (the terminating null character).

Each character stored in the string may occupy more than one byte. The encoding used to represent characters in a multibyte character string is locale-specific: it may be UTF-8, GB18030, EUC-JP, Shift-JIS, etc. For example, the char array `{'\xe4','\xbd','\xa0','\xe5','\xa5','\xbd','\0'}` is an NTMBS holding the string `"你好"` in UTF-8 multibyte encoding: the first three bytes encode the character 你, the next three bytes encode the character 好. The same string encoded in GB18030 is the char array `{'\xc4', '\xe3', '\xba', '\xc3', '\0'}`, where each of the two characters is encoded as a two-byte sequence.

In some multibyte encodings, any given multibyte character sequence may represent different characters depending on the previous byte sequences, known as "shift sequences". Such encodings are known as state-dependent: knowledge of the current shift state is required to interpret each character. An NTMBS is only valid if it begins and ends in the initial shift state: if a shift sequence was used, the corresponding unshift sequence has to be present before the terminating null character. Examples of such encodings are BOCU-1 and SCSU (http://www.unicode.org/reports/tr6) .

A multibyte character string is layout-compatible with null-terminated byte string (NTBS), that is, can be stored, copied, and examined using the same facilities, except for calculating the number of characters. If the correct locale is in effect, I/O functions also handle multibyte strings. Multibyte strings can be converted to and from wide strings using the following locale-dependent conversion functions:

Multibyte/wide character conversions

Defined in header <stdlib.h>	
mblen	returns the number of bytes in the next multibyte character (function)
mbtowc	converts the next multibyte character to wide character (function)
wctomb **wctomb_s** (C11)	converts a wide character to its multibyte representation (function)
mbstowcs **mbstowcs_s** (C11)	converts a narrow multibyte character string to wide string (function)
wcstombs **wcstombs_s** (C11)	converts a wide string to narrow multibyte character string (function)
Defined in header <wchar.h>	
mbsinit (C95)	checks if the mbstate_t object represents initial shift state (function)
btowc (C95)	widens a single-byte narrow character to wide character, if possible (function)
wctob (C95)	narrows a wide character to a single-byte narrow character, if possible (function)
mbrlen (C95)	returns the number of bytes in the next multibyte character, given state (function)
mbrtowc (C95)	converts the next multibyte character to wide character, given state (function)
wcrtomb (C95) **wcrtomb_s** (C11)	converts a wide character to its multibyte representation, given state (function)
mbsrtowcs (C95) **mbsrtowcs_s** (C11)	converts a narrow multibyte character string to wide string, given state (function)
wcsrtombs (C95) **wcsrtombs_s** (C11)	converts a wide string to narrow multibyte character string, given state (function)
Defined in header <uchar.h>	
mbrtoc16 (C11)	generates the next 16-bit wide character from a narrow multibyte string (function)
c16rtomb (C11)	converts a 16-bit wide character to narrow multibyte string (function)
mbrtoc32 (C11)	generates the next 32-bit wide character from a narrow multibyte string (function)
c32rtomb (C11)	converts a 32-bit wide character to narrow multibyte string (function)

Types

Defined in header <wchar.h>	
mbstate_t (C95)	conversion state information necessary to iterate multibyte character strings (class)
Defined in header <uchar.h>	
char16_t (C11)	16-bit wide character type

(typedef)

char32_t (C11)	32-bit wide character type (typedef)

Macros

Defined in header <limits.h>

MB_LEN_MAX	maximum number of bytes in a multibyte character, for any supported locale (macro constant)

Defined in header <stdlib.h>

MB_CUR_MAX	maximum number of bytes in a multibyte character, in the current locale (macro variable)

Defined in header <uchar.h>

__STDC_UTF_16__ (C11)	indicates that UTF-16 encoding is used by mbrtoc16 and c16rtomb (macro constant)
__STDC_UTF_32__ (C11)	indicates that UTF-32 encoding is used by mbrtoc32 and c32rtomb (macro constant)

References

- C11 standard (ISO/IEC 9899:2011):
  - 7.10 Sizes of integer types <limits.h> (p: 222)
  - 7.22 General utilities <stdlib.h> (p: 340-360)
  - 7.28 Unicode utilities <uchar.h> (p: 398-401)
  - 7.29 Extended multibyte and wide character utilities <wchar.h> (p: 402-446)
  - 7.31.12 General utilities <stdlib.h> (p: 456)
  - 7.31.16 Extended multibyte and wide character utilities <wchar.h> (p: 456)
  - K.3.6 General utilities <stdlib.h> (p: 604-614)
  - K.3.9 Extended multibyte and wide character utilities <wchar.h> (p: 627-651)
- C99 standard (ISO/IEC 9899:1999):
  - 7.10 Sizes of integer types <limits.h> (p: 203)
  - 7.20 General utilities <stdlib.h> (p: 306-324)
  - 7.24 Extended multibyte and wide character utilities <wchar.h> (p: 348-392)
  - 7.26.10 General utilities <stdlib.h> (p: 402)
  - 7.26.12 Extended multibyte and wide character utilities <wchar.h> (p: 402)
- C89/C90 standard (ISO/IEC 9899:1990):
  - 4.1.4 Limits <float.h> and <limits.h>
  - 4.10 GENERAL UTILITIES <stdlib.h>
  - 4.13.7 General utilities <stdlib.h>

See also

C++ documentation for **Null-terminated multibyte strings**