

The New C Standard (Excerpted material)

An Economic and Cultural Commentary

Derek M. Jones

derek@knosof.co.uk

3.7.2

multibyte character

multibyte character

sequence of one or more bytes representing a member of the extended character set of either the source or the execution environment

Commentary

Multibyte characters are a means of representing characters from those character sets that contain more members than can be represented in a byte.^[1] Somewhat confusingly the term *extended character set* is used by the C Standard to denote all supported characters, not just the extended characters.

A multibyte character is usually made up of a sequence of bytes that can be generated by pressing keys on the commonly available keyboards. There is not usually an obvious correspondence between the sequence of byte values and the numeric value of a member of the execution character set (such a correspondence does exist for a wide character), but there is an algorithm (often specified using a finite state machine) for converting them to this execution character set value.

Note: The use of the term *character* in the C Standard excludes multibyte characters, unless explicitly stated otherwise. This convention is also followed in the non-C Standard material in this book.

Common Implementations

The sequence of bytes in a multibyte character often has no relationship to what a developer types on a keyboard. For instance, at one end of the scale UTF-8 is unrelated to keystrokes. At the other end, many European keyboards have *dead keys* so that the single-byte characters *a-grave* or *i-circumflex* might be typed as *grave* followed by *a* or *circumflex* followed by *i*.

The most commonly used encoding methods include ISO 2022, EUC (Extended Unix Code), Big Five, Shift-JIS, and ISO 10646. Lunde^[1,2] covers East Asian characters and their representations and encodings in great detail.

NOTE The extended character set is a superset of the basic character set.

Commentary

The definition of the terms *basic character set* and *extended character set* implies that they are disjoint sets.

extended character set extended characters

multibyte character state-dependent encoding

ISO 2022

basic character set extended character set

References

1. K. Lunde. *Understanding Japanese Information Processing*. O'Reilly & Associates, Inc., 1993.
2. K. Lunde. *CJKV Information Processing*. O'Reilly & Associates, Inc., 1999.