D1885R8 Text encodings

Encodings, octets, bytes, wchar_t

D1885R8 Text encodings

Wikipedia

“The octet is a unit of digital information in computing and telecommunications that consists of eight bits.”

D1885R8 Text encodings

ISO 10646

3.22 encoding form
form that determines how each UCS code point for a UCS character is to be expressed as one or more code units used by the encoding form

D1885R8 Text encodings

ISO 10646

3.23
encoding scheme
scheme that specifies the serialization of the code units from the encoding form into octets

D1885R8 Text encodings

ISO 10646 section 11.5 “UTF-16”

The UTF-16 encoding scheme serializes a UTF-16 code unit sequence by ordering octets [little-endian or big-endian].
In the UTF-16 encoding scheme, the initial signature read as <FE FF> indicates [big-endian], and <FF FE> the reverse.
The signature is not part of the textual data.
In the absence of signature, the octet order of the UTF-16 encoding scheme is [big-endian].

D1885R8 Text encodings

RFC 2978 section 1.3
https://www.iana.org/assignments/character-sets/character-sets.xml

The term "charset" [...] is used here to refer to a method of converting a sequence of octets into a sequence of characters.

D1885R8 Text encodings

C++[lex.string] p10

String literal objects are initialized with the sequence of code unit values corresponding to the string-literal's sequence of s-char s (for a non-raw string literal) and r-char s (for a raw string literal) in order as follows: [...]

D1885R8 Text encodings

C++[lex.charset] p5

“Characters in a *character-literal* […] or in a *string-literal* are encoded as a sequence of one or more code units, as determined by the *encoding-prefix* (5.13.3, 5.13.5); this is termed the respective literal encoding.”

D1885R8 Text encodings

C++[intro.memory] p1

“The fundamental storage unit in the C ++ memory model is the byte. A byte is at least large enough to contain the ordinary literal encoding of any element of the basic literal character set (5.3) and the eight-bit code units of the Unicode UTF-8 encoding form and […]”

D1885R8 Text encodings

C++[basic.types.general] p4

“The object representation of an object of type T is the sequence of N unsigned char objects taken up by the object of type T, where N equals sizeof(T).”

D1885R8 Text encodings

iconv

```
size_t iconv(iconv_t cd,
             char **inbuf, size_t *inbytesleft,
             char **outbuf, size_t *outbytesleft);
```

The iconv() function converts a sequence of characters in one character encoding to a sequence of characters in another character encoding.