文字化け

(Mojibake)

by

Martinho Fernandes

UTF-8 (hex) : 4D617274696E686F204665726E616E646573

[maɾˈtiɲu]   [fəɾˈnandəʃ]

http://flamingdangerzone.com

15.04.2014 C-Base

Unicode

WARNING!

⚠ ← U+26A0
(NOT ASCII

NON-ASCII
CHARACTER PORN
AHEAD

OR

THIS WOULD NOT BE A TALK
ABOUT UNICODE 😉

U+1F609
(NOT ASCII)
(NOT BMP)

the C-word

A ?

the C-word

Á?

the C-word

-?

the C-word

Á?

the C-word

ł?

Á?

the C-word

한?

Á?

漢?

the C-word

한?

Á?

漢?

the C-word

한?

。?

Á?

漢?

the C-word

한?

ـط?ه?

Á?

漢?

the C-word

한?

هـ? ـه? ه?

Á?

漢?

the C-word

한?

ـه? ـهـ? هـ? ه?

Á?

漢?

the C-word

الله?

한?

هـ? ـهـ? ـه? ه?

U+D800?

Á?

漢?

the C-word

اللّه?

한?

ﻫ? ﻬ? ـﻪ? ه?

U+D800?

0x80?

Á?

漢?

the C-word

اللّٰه?

한?

ﻫ?ﻬ?ﻪ?ه?

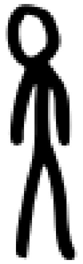
U+D800?

0x80?

Á?

漢?

the C-word

 ?

الله?

한?

ﻫ? ـﻬـ? ـﻪ? ه?

U+D800?

Á?

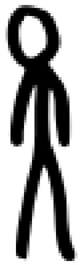
wchar_t?

0x80?

漢?

char?

the C-word

?

اللّٰه?

한?

ﻫ? ﻬ? ﻪ? ه?

U+D800?

wchar_t?

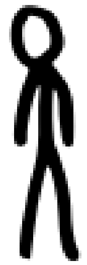
0x80?

Á?

漢?

char?

the C-word


?

[tʃa(ɹ)]
[kʰæ(ɹ)] ?

أللّٰه ?

한?

ـه? ـهـ? هـ? ه?

u"你好世界 😂"   char16_t[8]

UTF-16

4F60 597D 4E16 754C 0020 D83D DE02 0000

std::u16string

U"你好世界 😂" char32_t[7]

UTF-32

00004F60 0000597D 00004E16 0000754C 00000020 0001F602 00000000

std::u32string

u8"你好世界 😂"   char8_t

u8"你好世界 😂" ~~char8_t~~

u8"你好世界 😂" char[17]

UTF-8

E4 BD A0 E5 A5 BD E4 B8 96 E7 95 8C 20 F0 9F 98 82

std::u8string

u8"你好世界 😂"   char[17]

UTF-8

E4 BD A0 E5 A5 BD E4 B8 96 E7 95 8C 20 F0 9F 98 82

std::~~u8~~string

u8"你好世界 😂" char[17]

UTF-8

E4 BD A0 E5 A5 BD E4 B8 96 E7 95 8C 20 F0 9F 98 82

std::string

std::wstring_convert

std::wbuffer_convert

bytes ↔ "wide string"
"wide stream"

codecvt_utf8<char16_t>

codecvt_utf8<wchar_t>

codecvt_utf8<char32_t>

codecvt_utf16<char16_t>

codecvt_utf16<wchar_t>

codecvt_utf16<char32_t>

codecvt_utf8_utf16<char16_t>

codecvt_utf8_utf16<wchar_t>

UCS-2?

codecvt_utf8<char16_t>

codecvt_utf8<wchar_t>

codecvt_utf8<char32_t>

codecvt_utf16<char16_t>

codecvt_utf16<wchar_t>

codecvt_utf16<char32_t>

codecvt_utf8_utf16<char16_t>

codecvt_utf8_utf16<wchar_t>

<locale>

```
bool  isspace(charT c, const locale& loc);
charT toupper(charT c, const locale& loc);
```

<locale>

```
bool isspace(charT c, const locale& loc);

charT toupper(charT c, const locale& loc);
```

```
u8"ß"
  / \
C3   9F
```

<regex>

UTS #18

Not even Level 1

std::regex("...") matches u8"中"

ICU

Boost.Locale

SO answer this talk is based on:

http://stackoverflow.com/a/17106065

Recent Unicode support proposal that didn't make it:

http://www.open-std.org/jtc1/wg21/docs/papers/2013/n3572.html