Subject: Re: Will UPP support full UNICODE (21bits long codepoint)?
Posted by mirek on Sat, 15 Aug 2020 09:33:01 GMT
View Forum Message <> Reply to Message

Oblivion wrote on Sat, 15 August 2020 11:12Quote:int GraphemeLength(const char *s);
int GraphemeLength(const wchar *s);

That requires width tables.

How about generating the width tables with uniset scripts (used by xterm et al.)?

They can generate width tables for doublewidth, ambiguous width, unknown width and combining chars from the UnicodeData.txt (latest version).
They are quite comprehensive.

I already started using them here (For double width and combining chars, ATM):
https://github.com/ismail-yilmaz/upp-components/blob/master/ CtrlLib/Terminal/Cell.cpp

We can put these into a generic GetCharWidth(int c) function, then utilize them in GetGraphemeLength(const wchar *s)?

This might not be the definitve solution, but it is the battle-tested one out in the wild.

Only real downside is that the tables might need updating albeit ifrequently.

Best regards,
Oblivion

Just to make sure we are at the same page: I am not speaking about graphics width here, but a number of bytes (or words) that form a single grapheme ("combined character").

EDIT: Still not clear enough: Number of bytes of the first grapheme at s.

Mirek