Subject: Re: Choosing the best way to go full UNICODE
Posted by mirek on Wed, 31 May 2017 13:25:51 GMT
View Forum Message <> Reply to Message

cbpporter wrote on Wed, 31 May 2017 15:06OK, let's see it in action.

I still think that MString is a text book case of Unicode indexability fallacy, but maybe I'm wrong.

And if it works, maybe it is fine.

Please let me know if you need some help for the basic stuff. I can also review Unicode conformity of algorithms.

Do you want go full Utf8 minimal valid sequence validation and overlong prevention?

Code Points	First Byte	Second Byte	Third Byte	Fourth Byte
U+0000..U+007F	00..7F			
U+0080..U+07FF	C2..DF	80..BF		
U+0800..U+0FFF	E0	A0..BF	80..BF	
U+1000..U+CFFF	E1..EC	80..BF	80..BF	
U+D000..U+D7FF	ED	80..9F	80..BF	
U+E000..U+FFFF	EE..EF	80..BF	80..BF	
U+10000..U+3FFFF	F0	90..BF	80..BF	80..BF
U+40000..U+FFFFF	F1..F3	80..BF	80..BF	80..BF
U+100000..U+10FFFF	F4	80..8F	80..BF	80..BF

With the rest error escaped?

Not sure about that, but what I know for sure that I want to put decoding in single template (unlike current charset.cpp) so that it can be fixed easily...