Subject: Re: Building & using U++ without TheIDE
Posted by mirek on Sat, 15 Sep 2007 20:48:08 GMT
View Forum Message <> Reply to Message

sergei wrote on Sat, 15 September 2007 16:16BOM: http://en.wikipedia.org/wiki/Byte_Order_Mark

Ah, thanks, I did not knew that there is one for UTF-8...

So, if I understand you well, EF BB BF at the start of UTF-8 sequence should be ignored, right?

Anyway, this rather looks like file issue... Not sure it should be part of basic UTF-8 code?

Quote:
Another problem is that ToSystemCharset always returns String, which is castable to char*, and TSTR is castable to _TCHAR - which would always be what Win32 functions expect.

Ah, I see. You just did not undestood me. Look at PocketPC versio - that one returns WString instead of String. The result of these goes always directly to system calls (or, for FromSystemCharset, directly uses value returned by system call), therefore this is possible.

IMO, all really need to do to try playing with UNICODE/TCHAR is to change #ifdefs to use PocketPC (WinCE) version for UNICODE too.

Quote:
Is UTF-8 correctly converted in String <-> WString conversions?

I hope so, for basic plane and except the BOM. There is one extension though: Wrong UTF-8 sequences are interpreted byte by byte as characters 0xEExx (where xx is the byte).

The purpose is simple: This way, you can convert ANY input data to WString and back without loosing any information. This proved the absolute neccessity if you have the editor capable of handling multiple encodings in single file.

(That is why we named this UTF-8EE, as "Error Escape" or "placing to 0xEExx")

Quote:
And are other plane symbols also supported (symbols that take 4 bytes in UTF-16)?

No, unfortunately, other planes are not implemented yet.

Mirek