Subject: Re: 16 bits wchar
Posted by mirek on Wed, 03 Oct 2007 08:11:49 GMT
View Forum Message <> Reply to Message

cbpporter wrote on Wed, 03 October 2007 00:16luzr wrote on Mon, 01 October 2007 14:28
I guess fixing Utf8 routines to provide UTF16 surrogate support (for now) is a good idea.

Great! On the side note though, I am extending Utf8 methods to handle 4 byte long encodings, not UTF-16 surrogate pairs (which are illegal in UTF-8).

Well, so what is the result then? WString is now 16-bit. Utf8 conversions are basically String<->WString (ok, also char * <-> WString).

Quote:
That could be an acceptable compromise. But a few processing functions couldn't hurt when you really want to process that string in place.

Which exactly?

Quote:
I think you should keep UTF-16 as default for Win32 and UTF-32 as default for Linux. Win32 and .NET both use UTF-16 (with surrogates - Win98 doesn't support surrogates, but the rest do), so I think the future of character encoding for GUI purposes is pretty well defined.

That is why it is 16bit now. But if you really need the solution for ucs-4, 32bit character and conversions is the only option.

Quote:

```
String  ToUtf8(wchar code);
String  ToUtf8(const wchar *s, int len);
String  ToUtf8(const wchar *s);
String  ToUtf8(const WString& w);

WString FromUtf8(const char *_s, int len);
WString FromUtf8(const char *_s);
WString FromUtf8(const String& s);

bool utf8check(const char *_s, int len);

int utf8len(const char *s, int len);
```

```
int utf8len(const char *s);
int lenAsUtf8(const wchar *s, int len);
int lenAsUtf8(const wchar *s);

bool    CheckUtf8(const String& src);
```

Quote:
2. Your GetUtf8 method is quite straightforward, but I'm afraid it does not decode values correctly.

Here is a pseudo code of what you do:

```
if(code <= 0x7F)
   compute 1 byte value
else if (code <= 0xDF)
   compute 2 byte value
else if (code <= 0xEF)
   compute 3 byte value
else if (...)
   pretty much just read them and return "space"
```

Ops, you are right, something is really missing in Stream. Anyway, GetUtf8 in Stream is quite auxiliary (and maybe wrong) addition. The real meat is in Charset.h/.cpp.

Mirek
