Subject: Re: 16 bits wchar
Posted by mirek on Tue, 25 Sep 2007 21:18:45 GMT
View Forum Message <> Reply to Message

Quote:
GetLength should return the length in code units and GgtCount the number of real characters, so code points.

I am afraid you expect too much. GetLength returns exactly the same number as GetCount, two names in this case are there because of each fits better for different scenario (same thing as 0 and '\0').

Rather than thinking in terms of UTF-8 / UTF-16... String is just an array of bytes, WString array of 16bit words. Not much more logic there, except that conversions between two can be performed - in conversions there is one and only encoding logic.

Quote:
I also started researching the exact encoding methods of UTF and I will add full Unicode support to strings. It will be for personal use, but if anybody is interested I will post my results. Right now I'm trying to find and efficient way to index multichar strings. I think I will have to use iterators instead.

Actually, it is not that I am not worried here. Anyway, I think that the only reasonable approach is perhaps to change wchar to 32-bit characters OR introduce LString.

The problem is that in that case you immediately have to perform conversions for all Win32 system calls... that is why I have concluded that it is not worth the trouble for now. (E.g. RTL clearly is the priority).

Anyway, any research in this area is welcome. And perhaps you could fix UTF-8 functions to support UTF-16 (so far, everything >0xffff is basically ignored).

Mirek