Subject: Re: 16 bits wchar
Posted by mirek on Fri, 12 Oct 2007 15:03:03 GMT
View Forum Message <> Reply to Message

cbpporter wrote on Fri, 12 October 2007 07:54
luzr wrote on Fri, 12 October 2007 11:59
Also interesting question: While longer UTF-8 sequences are invalid, would not be actually a good idea to accept them as a form of error-escapement? I can imagine a couple of scenarious where this might be very useful... E.g. what are we supposed to to with invalid UCS-4 values after all?

Yes, that would also be a good alternative. I choose the EExx encoding out of two reasons:

Actually, I would keep EExx for ill-formed utf8 anyway. What I was up to was rather the fact that UTF-8 represents a sort of huffman encoding.

In practice, there is a lot of cases where you have store a set of offsets or indicies efficiently, which are "small" (e.g. lower than 128) in most case, but in exceptional cases they can be larger.

Using "full" UTF-8 would provide a nice compression algorithm here...

(Note that such use is completely unrelated to UNICODE, but why not to reuse the existing code?.

Mirek