

Re: wstring to ostream

Source: <http://www.tech-archive.net/Archive/VC/microsoft.public.vc.stl/2006-12/msg00082.html>

- *From:* MrAsm <invalid@xxxxxxxxxxxx>
 - *Date:* Thu, 21 Dec 2006 07:20:50 GMT
-

On Thu, 21 Dec 2006 06:49:20 +0100, Howie Meier <howieh@xxxxxxx> wrote:

I thought only wstring can store UTF-8. How can i convert the UTF-8 ?
Have you got an example or link etc. ?

Thanks Mr.Asm.

Howie

There are different encodings for Unicode characters; UTF-8 and UTF-16 are two examples of Unicode character encodings.

In UTF-8 encoding, a Unicode character can be stored in 1 or 2 or 3 or even 4 bytes. The base unit in UTF-8 are 8 bits (1 byte). The UTF-8 is so a pure sequence of bytes (back-compatible with English ASCII characters, i.e. e.g. no italian vowels with accents like è or é.)

In UTF-16 encoding, a Unicode character can be stored in one or two "code unit"; a code unit in UTF-16 is always 16 bits (2 bytes). A lot of characters need only one UTF-16 code unit (i.e. 2 bytes), but there are also some characters (Chinese characters, some musical symbols, characters from ancient alphabets, etc.) that can have so called "surrogates", and they need two code units (2 x 16 bits).

So, while a UTF-8 encoded string is a sequence of chars (or CHARs in Win32 SDK), a UTF-16 encoded string is a sequence of wchar_t's (16 bits, or WCHARs in Win32 SDK).

So, you can store a UTF-8 encoded string in std::string (or even std::vector< char >), and a UTF-16 string in std::wstring (or even std::vector< wchar_t >).

To convert from UTF-8 to UTF-16, you can use the ::MultiByteToWideChar() Win32 function; to convert from UTF-16 to UTF-8, you can use the ::WideCharToMultiByte() Win32 function.

Re: wstring to ostream

Use CP_UTF8 as code page (this is one of the parameters required by these functions).

If you want more details, you could also search on microsoft.public.vc.mfc, when Unicode has been discussed recently, too.

I would also suggest the following links:

<<http://en.wikipedia.org/wiki/UTF-8>>

<http://unicode.org/unicode/faq/utf_bom.html#UTF8>

<<http://www.unicode.org/unicode/faq/>>

Hope this helps.

Mr.Asm

.