

Writing UTF-8 file under Windows

Source: <http://coding.derkeiler.com/Archive/Perl/comp.lang.perl.misc/2007-01/msg00223.html>

- *From:* t_lawetta@xxxxxxxxxx
 - *Date:* 5 Jan 2007 05:37:16 -0800
-

Happy New Year,

Whatever I try to write a UTF-8 file, I always end up with UTF-16LE with the "FF FE" BOM at the beginning and 2 bytes per character.

I am reading strings from an external resource and try to write to files.

```
my $string_with_special_chars = "Château Müller\nGarçon";
open F, ">:utf8", "test.txt";
print F $string_with_special_chars;
close F;
```

Tried it both on Linux (Perl 5.8.6) and Windows (Perl 5.8.7).
(In case you cannot see it: The string contains the chars with the corresponding HTML entities acirc, uuml and ccedil.

Opening test.txt with my editor (Ultra-Edit) shows me the correct string, but in hex view I see the "FF FE" BOM and it shows 2 bytes per character, e.g. 0x43 0x00 for the 'C' and 0xE7 0x00 for the ccedil.

Normally I am reading data via LDAP, so 'use utf8' is not required. If I add it here, I get:
Malformed UTF-8 character (unexpected non-continuation byte 0x74, immediately after start byte 0xe2) at ./test.pl line 4.

I tried to make sure my input strings are correctly decoded etc., but no way.

As long as my strings stay within 7-bit ASCII it is fine, but after that Perl always things it has to write a BOM and decode in a 2-byte format.

Using Encode to write utf-8 results in a double encoding or at least some unreadable chars.

Where does the BOM come from?
Why does Perl add it?
Doesn't Perl write UTF-8 by default?

Writing UTF-8 file under Windows

Thank you for any hints. The issue cost me days already and yes, I have read a lot about Perl and Unicode.

Tony

.