DCMTK - Bug #386

Check whether the VR-Scanner can manage UTF-8 and other MBCS

2011-02-17 00:00 - Jörg Riesmeier

Status:	New	**Start date:**	
Priority:	Normal	**Due date:**	
Assignee:		**% Done:**	0%
Category:	Library	**Estimated time:**	0:00 hour
Target version:			
Module:	dcmdata	**Compiler:**	
Operating System:			

Description

... noticed while checking UTF-8 records with DCMCHECK 3.0 pre

Excerpt from 'dcmdata/libsrc/vrscan.l'

```
default_charset_without_control_chars [\041-\133\135-\176][\040-\133\135-\176]*
charset_without_control_chars         [\040-\133\135-\176\240-\377\033]+
charset_with_control_chars            [\040-\176\240-\377\012\014\015\033]+
```

History

#1 - 2013-02-13 17:27 - Jörg Riesmeier

- Category set to Library

- Target version set to 3.6.2

#2 - 2013-03-19 09:50 - Uli Schlachter

For UTF-8, Wikipedia shows which byte values are valid: http://en.wikipedia.org/wiki/UTF-8#Codepage_layout

All together, UTF-8 encoded text can contain bytes in the range 0x00-0xbf and 0xc2-0xf4. In other words, invalid bytes values would be 0xc0-0xc1 and 0xf5-0xff. Everything but these 13 values can appear in UTF-8 encoded text. (The original UTF-8 before it was limited to U+10ffff can even contain every byte except for 0xc0, 0xc1, 0xfe and 0xff (which are used for UTF-16's byte order mark)).

However, since UTF-8 is compatible with ASCII, it should be safe to continue to forbid values 0x00-0x1f (various control characters). This means we are left with [\040-\176\200-\277\302-\377\012\014\015\033]+ for charset_with_control_chars. (Why is \033 "Escape" allowed in the current vrscanner? Why is \134 "\" not allowed in without_control_chars? Should \177 "DEL" be allowed?)

The difference to the current version is that \200-\237 are now allowed, too. These are 0x80-0x9f and are continuation bytes in UTF-8. These values can appear in almost any script... Also, \300 and \301 get forbidden. This is likely a bad idea, because other encodings than UTF-8 use them (?).

#3 - 2013-04-28 23:24 - Andrew Chiw

- Subject changed from Prüfen inwieweit der VR-Scanner mit UTF-8 und anderen MBCS zurecht kommt to Check whether the VR-Scanner can manage UTF-8 and other MBCS

#4 - 2017-03-24 12:41 - Marco Eichelberg

- Target version changed from 3.6.2 to 3.6.3

#5 - 2017-03-24 12:51 - Marco Eichelberg

- Priority changed from High to Normal

#6 - 2017-12-07 11:58 - Marco Eichelberg

- Target version changed from 3.6.3 to 3.6.6

#7 - 2020-05-25 13:28 - Michael Onken

- Target version deleted (3.6.6)