Förändrad teckenkodning i TIF

Innehållsförteckning

1 Inledning

1.1 Bakgrund

Taric Internet Fildistibution (TIF) har sedan starten för drygt tio år sedan använt den västeuropeiska teckenkodningen ISO-8859-1 för alla distributionsfiler som läggs ut på vår FTP-server. Vi avser att ändra kodningen till den mer moderna och internationellt anpassade teckenkodningen UTF-8.

Taric3 som infördes av Tullverket i mars 2011 använder UTF-8 i grunden. Det ger användaren möjlighet att presentera helt korrekta beskrivningstexter med ovanliga tecken och symboler som inte finns tillgängliga i ISO-8859-1.

Förändringen gör det möjligt för den som använder TIF att kunna få helt korrekta och lättförståeliga beskrivningstexter med olika typer av ovanliga tecken och symboler som används i Taric. Sedan starten av TIF har det funnits problem med att översätta vissa ovanliga tecken. Det kan hanteras om man använder UTF-8. Det har även framkommit önskemål från användare om att få UTF-8 teckenkodning i filerna. Dessutom internationaliseras lösningen och säkerställs för framtiden på ett bättre sätt.

1.2 Sammanfattning

Tullverket kommer att förändra teckenkodningen för TIF:s distributionsfiler i både ”flat” och XML-format från dagens ISO-8859-1 till UTF-8 format. Införandet sker från och med måndag den 15 april 2013.

De som använder TIF och nyttjar posttyper med beskrivningstexter kan välja att anpassa sitt användande av dem till UTF-8 och få fördelarna som det innebär eller att inte göra någon åtgärd alls men istället kanske få problem med visningen av vissa vanliga tecken som t.ex. å, ä, ö.

Om man använder filer utan beskrivningstexter behövs inget göras då allt i grunden kommer fungera som tidigare.

2 Förändringsbeskrivning

2.1 Allmänt

Teckenuppsättningarna mellan ISO-8859-1 och UTF-8 stämmer delvis överens med i UTF-8. Representationen ser lika ut om man tittar i en fil. Skillnaden är att i UTF-8 finns möjlighet att lagra all världens tecken och symboler. Det sker med genom att vissa tecken representeras med 2-4 databyte istället för 1 databyte.

Om man öppnar en fil i UTF-8 format i en textredigerare eller annat program som inte stödjer detta format så kan man exempelvis se flerdatabyte tecknen nedan.

å = Ã¥
ä = Ã¤
ö = Ã¶
é = Ã©
µ = Î¼
α = Î±
± = Â±
ω = Ï‰
• = â€¢

Läs mer om UTF-8 på t.ex. http://sv.wikipedia.org/wiki/UTF-8

2.2 Redigeringstecken i texterna blir oförändrade

I beskrivningstexterna förekommer vissa redigeringstecken som användes innan UTF-8 fanns i Taric. Dessa kommer tillsvidare att finnas kvar fast de inte längre behövs.

!1!	Radbrytning
<P>	Radbrytning (skall enligt uppgift ersätta !1!)
|	Hårt mellanslag
$	Efterföljande tecken skall vara upphöjt (superscript)
@	Efterföljande tecken skall vara nedsänkt (subscript)
!%!	Per tusen
!X!	Multiplikation
!o!	Grader
!>=!	Större än eller lika med

2.3 Några förbättringar med UTF-8

- Kan hantera alla världens tecken och symboler.
- Är mer vanligt som standarduppsättning i operativsystem och nya program.
- De översättningsproblem som finns idag och som resulterar i ett ’?’ i distributionsfilen kommer att försvinna.

2.3.1 Exempel på beskrivningstexter

Här kan ni se några exempel på hur varukodsbeskrivningar ser ut i filerna respektive vid presentation.

Varukod 3926909755

ISO-8859-1

----Platta produkter av polyeten, perforerade i motstående riktningar, med en tjocklek av minst 600|?m men högst 1|200|?m och en vikt av minst 21|g/m$2 men högst 42|g/m$2

Ger nedanstående vid presentation.

----Platta produkter av polyeten, perforerade i motstående riktningar, med en tjocklek av minst 600 ?m men högst 1 200 ?m och en vikt av minst 21 g/m^2 men högst 42 g/m^2

UTF-8

----Platta produkter av polyeten, perforerade i motstÃ¥ende riktningar, med en tjocklek av minst 600|Î¼m men hÃ¶gst 1|200|Î¼m och en vikt av minst 21|g/m$2 men hÃ¶gst 42|g/m$2

Ger nedanstående vid presentation.

----Platta produkter av polyeten, perforerade i motstående riktningar, med en tjocklek av minst 600 µm men högst 1 200 µm och en vikt av minst 21 g/m^2 men högst 42 g/m^2

Varukod 2907290070

ISO-8859-1

---2,2´,2´´,6,6´,6´´-Hexa-_tert_-butyl-_?,?´,?´´_-(mesitylen-2,4,6-triyl)tri- _p_-kresol (CAS RN 1709-70-2)

Ger nedanstående vid presentation.

---2,2′,2″,6,6′,6″-Hexa-_tert_-butyl-_?,?′,?″_-(mesitylen-2,4,6-triyl)tri- _p_-kresol (CAS RN 1709-70-2)

UTF-8

---2,2â~@²,2â~@³,6,6â~@²,6â~@³-Hexa-_tert_-butyl-_Î±,Î±â~@²,Î±â~@³_-(mesitylen-2,4,6-triyl)tri- _p_-kresol (CAS RN 1709-70-2)

Ger nedanstående vid presentation.

---2,2′,2″,6,6′,6″-Hexa-_tert_-butyl-_α,α′,α″_-(mesitylen-2,4,6-triyl)tri- _p_-kresol (CAS RN 1709-70-2)

2.4 Förändrade posttyper och distributionsfiler

Posttyperna D, E, F, K, M, O, R, Q och T påverkas. I filerna ändras bara själva beskrivningstexterna, övrig hantering av filerna blir oförändrad.
Filerna i ”flat” respektive XML- format ändras och det gäller för både total samt differensfilerna.

(#### = serienummer)

Posttyp D - Varuslagstexter
Fältet: LONG_DESCR

Filer: ####_KD_EN.tot, ####_KD_EN.totxml, ####_KD_SV.tot, ####_KD_SV.totxml, ####_DD_EN.dif , ####_DD_EN.difxml, ####_DD_SV.dif, ####_DD_SV.difxml

Posttyp E - Tilläggskodtexter
Fältet: LONG_DESCR

Filer: ####_KE_EN.tot, ####_KE_EN.totxml, ####_KE_SV.tot, ####_KE_SV.totxml, ####_DE_EN.dif , ####_DE_EN.difxml, ####_DE_SV.dif, ####_DE_SV.difxml

Posttyp F - Fotnotstexter
Fältet: LONG_DESCR

Filer: ####_KF_EN.tot, ####_KF_EN.totxml, ####_KF_SV.tot, ####_KF_SV.totxml, ####_DF_EN.dif , ####_DF_EN.difxml, ####_DF_SV.dif, ####_DF_SV.difxml

Posttyp K – Procentsatser till jordbrukskomponenter (meursing)
Fältet: SHORT_DESCR

Filer: ####_KK.tot, ####_KK.totxml, ####_DK.dif , ####_DK.difxml

Posttyp M - Exportbidragsnomeklatur
Fältet: LONG_DESCR

Filer: ####_KM_EN.tot, ####_KM_EN.totxml, ####_KM_SV.tot, ####_KM_SV.totxml, ####_DM_EN.dif , ####_DM_EN.difxml, ####_DM_SV.dif, ####_DM_SV.difxml

Posttyp O – Nationella skatter och avgifter
Fältet: LONG_DESCR

Filer: ####_KO.tot, ####_KO.totxml, ####_DO.dif , ####_DO.difxml

Posttyp Q – Nationella skatter och avgifter, fotnotstexter

Fältet: LONG_DESCR

Filer: ####_KQ.tot, ####_KQ.totxml, ####_DQ.dif , ####_DQ.difxml

Posttyp R - Kodförteckning
Fältet: SHORT_DESCR

Filer: ####_KR_EN.tot, ####_KR_EN.totxml, ####_KR_SV.tot, ####_KR_SV.totxml, ####_DR_EN.dif , ####_DR_EN.difxml, ####_DR_SV.dif, ####_DR_SV.difxml

Posttyp T - Växelkurser
Fältet: EXCH_NAME
Fältet: CTRY_TEXT

Filer: ####_KT.tot, ####_KT.totxml, ####_DT.dif , ####_DT.difxml

2.5 Införandeplan av UTF-8

2.5.1 Driftsättning

Driftsättning planeras till måndagen den 15 april 2013.

2.5.2 Arbetsgång vid driftsättningen

Under måndag eftermiddag på driftsättningsdagen kommer nya totalfiler med UTF-8 formatering att finnas klara för nedladdning. De kommer att ha samma serienummer som föregående körning föregående fredag. Senare vid vanlig tid ca: 22.00 kommer nya differensfiler i UTF-8 format läggas ut. De innehåller då bara förändringar från föregående körning och har sekvensnumret efter total-filerna som skapats under dagen.

Efter driftsättningen kommer total och diff-filer endast att finnas i UTF-8 format och läggas ut enligt tidigare rutin dvs. total-filer var 30:e körning samt diff-filer varje vardag.

2.6 Testfiler i UTF-8 format

Testfiler finns på webben http://distr.tullverket.se/taric/ och kan även nås via vår FTP server. Notera att filerna inte är PGP- signerade, behöver bara packas upp för att kunna läsas.
(Se testlänkarna längst ner på sidan)

3 Åtgärder för TIF - användare

3.1 Rekommenderade åtgärder

Vad ni behöver göra beror naturligtvis på hur datat från TIF används och hur lösningen har implementerats. Några punkter att tänka kan vara nedanstående.

- Testa ert verksamhetssystem eller användandet av TIF-filerna med framtagna UTF-8 testfiler, så att ni kan verifiera att textbeskrivningarna visas korrekt.
- Göra eventuella anpassningar av ert system eller själva inläsningen av filerna, så att ni får en korrekt presentation av texterna.
- Om ni går över till att lagra UTF-8 formaterat data i ert verksamhetssystem behöver ni rensa ert gamla TIF-data som lagras i ISO-8859-1 och ladda in nya total-filer vid driftsättningen.
- Vill ni behålla ISO-8859-1 i sitt system kan ni välja att konvertera UTF-8 till ISO-8859-1 vid t.ex. inläsningen av datat, men då har ni kvar samma översättningsproblem som finns idag.
- Ta fram en plan för ert införande vid själva driftsättningen.

3.2 Om ingen anpassning görs

Allt kommer i grunden fungera som tidigare vilket gör att det finns en möjlighet att ni inte behöver införa eventuella anpassningar på själva driftsättningsdagen. Om ni väljer att inte göra någonting alls får ni kanske problem med visning av t.ex. (ääö) i ert system. Om man inte rensar sitt TIF data och fortsätter att läsa in de dagliga differensfilerna kommer alla uppdaterade texter allteftersom bli i UTF-8 format och man får då en blandning av de två teckenkodningarna. Men det går naturligtvis alltid att rensa sitt data och ladda in nya totalfiler när man vill.