# Changed character encoding in TIF

## Contents

# 1 Introduction

## 1.1 Background

Taric Internet File Distribution (TIF) has since the beginning over 10 years ago used the Western European character encoding ISO-8859-1 in all the files transferred to our distribution server.

Taric 3, that was introduced by Swedish Customs in March 2011, is basically using UTF-8 which gives the user possibility to present completely accurate description files containing uncommon characters and symbols not available in ISO-8859-1.

The change makes it possible for the Taric users to obtain entirely accurate and easy-to-understand description files containing different types of unusual characters and symbols used in Taric. Since the beginning of TIF there have been problems to translate some unusual characters. This can be handled with UTF-8. There have also been requests from TIF users to get UTF-8 encoding in the files. Additionally, the solution will be internationalized and secured to the future in a better way.

## 1.2 Summary

Swedish Customs is about to change the character encoding in the TIF distribution files, both “flat files” and XML format, from the current format ISO-8859-1 to UTF-8. The deployment is scheduled on Monday 15 April 2013.

If you use TIF and use record types including description texts, you can choose either to adjust your use of them to UTF-8 and get all the benefits it brings or to take no action at all but instead having potential problems with the display of some common characters as for example å, ä, ö.

If you use files without description texts you don’t need to do anything at all because everything is basically going to operate as before.

# 2 Change up description

## 2.1 General

The character sets in ISO-8859-1 is partly consistent with UTF-8. The representation looks the same if you are looking in a file. The difference is that in UTF-8 it is possible to store all the characters and symbols in the world. This is done by representing certain characters with 2-4 bytes instead of one.

If you open a file in UTF-8 format in a text editor or other program that does not support the format, you can for example see the multi-byte characters shown below.

å = Ã¥

ä = Ã¤

ö = Ã¶

é = Ã©

μ = Î¼

α = Î±

± = Â±

ω = Ï‰

• = â€¢

You can read more about UTF-8 for example on this web page:
http://sv.wikipedia.org/wiki/UTF-8

## 2.2 Editing characters in the texts remain unchanged

In the description texts there are some editing characters that were used before UTF-8 was represented in Taric. They will remain for the time being although they are no longer needed.

| | |
|---|---|
| !1! | Newline |
| <P> | Newline (will replace !1!) |
| \| | Non-breaking space |
| $ | The following character shall be a superscript |
| @ | The following character shall be a subscript |
| !%! | Per mille (thousand) |
| !X! | Multiplication |
| !o! | Degrees |
| !>=! | Greater than or equal to |

## 2.3 Some improvements in UTF-8

- Can handle all the characters and symbols in the world.
- More common as the standard set in operating systems and new programs.

- The existing translation problems today, giving a ‘?’ in the distribution files, will disappear.

### 2.3.1 Examples of description texts

Below you can see some examples of how goods descriptions are shown in the files and at presentation.

**Commodity code 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

Provides the following in the 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

Provides the following in the 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$^{2}$

**Commodity code 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)

Provides the following in the 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)

Provides the following in the 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 Changed record types and distribution files

The record types D, E, F, K, M, O, R, Q and T will be affected. Only the description texts are changed in the files, all other operations will remain unchanged. The files in “flat” and XML format will be changed and this applies both to the total files and the difference files.

*(#### = serial number)*

### Record type D – Commodity code texts

Field: LONG_DESCR

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

### Record type E – Additional code texts

Field: LONG_DESCR

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

### Record type F – Footnote texts

Field: LONG_DESCR

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

### Record type K – Percentage for agricultural components (meursing)

Field: SHORT_DESCR

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

### Record type M – Export refund nomenclature

Field: LONG_DESCR

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

### Record type O – National taxes and duties

Field: LONG_DESCR

Filers ####_KO.tot, ####_KO.totxml, ####_DO.dif , ####_DO.difxml

### Record type Q – National taxes and duties, footnote texts

Field: LONG_DESCR

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

**Record type R – Code list**
Field: SHORT_DESCR

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

**Record type T – Exchange rates**
Field: EXCH_NAME
Fieldt: CTRY_TEXT

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

## 2.5 Implementation plan for UTF-8

### 2.5.1 Deployment

The deployment is scheduled on Monday 15 April 2013.

### 2.5.2 Workflow for deployment

On Monday afternoon on the deployment day, new total files containing UTF-8 formatting will be ready to download. They will have the same serial number as those from the previous run which was made the Friday before. Later at the same time as usual, about 10.00 PM, new difference files containing UTF-8 formatting will be transferred to the distribution server. They will only include changes from the previous run and have the sequence number following the one of the total files produced earlier the same day.

After the deployment, total and difference files will only exist in UTF-8 format. They will be transferred in the same way as before, namely total files every 30 days and difference files every weekday.

## 2.6 Test files in UTF-8 format

Test files are available on our website http://distr.tullverket.se/taric/ . They can also be accessed via our FTP server. Please note that the files aren’t PGP signed. You just have to unpack the files to be able to read them. (Se test links at the bottom of the page.)

# 3 Actions for the TIF user

## 3.1 Recommended actions

Naturally, what you need to do depends on how you use the TIF data and how your application has been implemented. Some points to consider may be those below:

- Test your business system or use of the TIF files with developed UTF-8 test files in order to verify that the text descriptions are shown correctly.
- Make any necessary adjustments in your system or the loading of the files in order to get a correct presentation of the texts.
- If you are going to store UTF-8 formatted data in your business system, you have to clear the old TIF data stored in ISO-8859-1 format and then load new total files at the time of deployment.
- If you want to keep ISO-8859-1 in your system you can choose to convert UTF-8 to ISO-8859-1, for example when loading, but you will still have the same translation problems as before.
- Make a plan for the introduction of your adjustments at the day of deployment.

## 3.2 If no adjustments are made

Basically, everything will work as before which means there is a possibility you don’t need to introduce any adjustments on the day of deployment. If you choose to do nothing, you may get a problem with displaying for example å, ä, ö in your system. If you don’t clear your TIF data and continue to load the daily difference files, all updated texts will be in UTF-8 format and you will get a mix of the two encodings. Of course, it is possible to clean your data at any time and load new total files whenever you want.