

ISO-8859 to UTF8 Converter

Skript, welches vom mir erstellt wurde, um alte Bullnix Pages (Post-Wiki) automatisiert in das neue UTF-8 Fileformat zu konvertieren! *Dies ist besonders wichtig, weil ansonsten Sonderzeichen wie, "ä", "ö", "ü" etc.. nicht mehr korrekt dargestellt werden können!*

Zu beachten: Das Skript, macht vor dem Konvertieren, jeweils ein Backup der original Dateien. *.backup! Falls es also Probleme gibt, enthalten jene Files den Inhalt der originalen Files!

Skript Sourcecode

Filename: **convert_ISO-8859_to_UTF8.sh**

```
#!/bin/bash

cd /data/webhome/bullnix-int.post.ch/pages/foswiki/data/ #Change to YOUR
folder, where the files are!!

for file_to_convert in $( find . -type f ); do

    file_encoding=$(file -i "$file_to_convert" | sed
"s/. *charset=\(.*\)$/\1/")
    if [ "${file_encoding}" = "iso-8859-1" ] || [ "${file_encoding}" =
"iso-8859-2" ];
    then
        cp -v "$file_to_convert" "${file_to_convert}.backup"
        iconv -f ISO-8859-1 -t UTF-8 "${file_to_convert}.backup" >
"$file_to_convert"
    else
        echo "File: ${file_to_convert} is already in 'utf8'
encoded!"
    fi
done
```

Um zu überprüfen, ob die Konvertierung auch für alle Files erfolgreich durchgelaufen ist, kann folgender Prompt gebraucht werden:

```
# cd /data/webhome/bullnix-int.post.ch/pages/foswiki/data/
# find . -name "*.txt" | while read F; do cat $F | hexdump -C | cut -
b11-60 | egrep -q "(c4|d6|dc|e4|e7|e8|e9|ea|f4|fc|f6)"; if [ $? -eq 0 ];
then echo "$F : ISO"; fi; done
```

Falls dieser Befehl nun Fehler ausgibt, kann folgendes Skript verwendet werden um diese zu korrigieren.

Filename: `fix_convertig_errors.sh`

```
#!/bin/bash

cd /data/webhome/bullnix-int.post.ch/pages/foswiki/data/

for file_to_fix in $( find . -name "*.*)" ); do

    #Detect ISO encodings in files via hexdump:
    cat $file_to_fix | hexdump -C | cut -b11-60 | egrep -q
    "(c4|d6|dc|e4|e7|e8|e9|ea|f4|fc|f6)";
    if [ $? -eq 0 ]; then

        echo "$file_to_fix : ISO";
        iconv -f ISO-8859-15 -t UTF-8 $file_to_fix >
temp_Convert.txt; cp temp_Convert.txt $file_to_fix

        sed -ri 's/\\xFC/ü/g' $file_to_fix
        sed -ri 's/\\xE4/ä/g' $file_to_fix
        sed -ri 's/\\xF6/ö/g' $file_to_fix

        sed -ri 's/\\xC4/Ä/g' $file_to_fix
        sed -ri 's/\\xDC/Ü/g' $file_to_fix
        sed -ri 's/\\xD6/Ö/g' $file_to_fix

        sed -ri 's/\\1E9E/□/g' $file_to_fix
        sed -ri 's/\\xDF/ß/g' $file_to_fix

        sed -ri 's/\\xE9/é/g' $file_to_fix
        sed -ri 's/\\xE8/è/g' $file_to_fix
        sed -ri 's/\\xEA/ê/g' $file_to_fix

        sed -ri 's/\\xE7/ç/g' $file_to_fix
        sed -ri 's/\\xC7/Ç/g' $file_to_fix

        #iconv -f ISO-8859-15 -t UTF-8 $file_to_fix >
temp_Convert.txt; cp temp_Convert.txt $file_to_fix
    fi;
done
```

Last update: **2017/09/11 11:37**