

UTF-8 Kodierung sicherstellen

== Was wir wollen ==

Die Anforderung ist einfach. Wir wollen einen UTF-8 String erhalten - egal welche Zeichenkodierung die Quelle hatte.

Entweder der String ist korrekt und sieht folgendermaßen aus:

Quellcode

```
1. $string_correct = "umlaute läuten die wende ein.";
```

Oder er ist falsch und sieht zum Beispiel so aus:

Quellcode

```
1. $string_wrong = "umlaute l•uten die wende ein.";
```

== Anwenden der Funktion utf8_encode() auf fehlerhafte Kodierung ==

Wenn wir wissen, dass die Eingabe fehlerhaft kodiert ist, dann fällt die Wahl einfach. Wir wenden die Funktion utf8_encode() also nur auf den falschen String an:

Quellcode

```
1. $string_wrong = "umlaute l•uten die wende ein.";
2. echo utf8_encode($string_wrong);
3. //Ausgabe = umlaute läuten die wende ein.
```

== Anwenden der Funktion utf8_encode() auf korrekte Kodierung ==

Doch was geschieht eigentlich wenn wir die Funktion utf8_encode() auf eine Zeichenkette anwenden, die bereits in UTF-8 vorliegt? Man könnte erwarten, dass wir weiterhin einen gültigen UTF-8 String erhalten, doch leider ist dem nicht so.

Quellcode

```
1. $string_correct = 'umlaute läuten die wende ein.';
2. echo utf8_encode($string_correct);
3. //Ausgabe = umlaute lÃuten die wende ein.
```

== Erkennen ob UTF-8 verwendet wird ==

Das Problem beschränkt sich also darauf zu erkennen ob der String in UTF-8 vorliegt. Dankenswerterweise liefert uns das PHP Manual die Lösung dazu und sagt uns welche Bit die Funktion utf8 selbst manipuliert.

easy-coding.de/Attachment/490/...883c82d04d57de50aa2daf6fd

Wir erstellen also die folgende Funktion um zu erkennen ob es sich bei einem String um einen UTF-8 String handelt - die Lösung wurde bereits im Jahre 2004 von "bmorel at ssi dot fr" gepostet.

Quellcode

```
1. function seems_utf8($Str) {
2.   for ($i=0; $i<strlen($Str); $i++) {
3.     if (ord($Str[$i]) < 0x80) continue; # 0bbbbbbb
4.     else if ((ord($Str[$i]) & 0xE0) == 0xC0) $n=1; # 110bbbbbb
5.     else if ((ord($Str[$i]) & 0xF0) == 0xE0) $n=2; # 1110bbbb
6.     else if ((ord($Str[$i]) & 0xF8) == 0xF0) $n=3; # 11110bbb
7.     else if ((ord($Str[$i]) & 0xFC) == 0xF8) $n=4; # 111110bb
8.     else if ((ord($Str[$i]) & 0xFE) == 0xFC) $n=5; # 1111110b
```

```

9. else return false; // Does not match any model
10. for ($j=0; $j<$n; $j++) {
12. // n bytes matching 10bbbbbb follow ?
13. if ((++$i == strlen($Str)) || ((ord($Str[$i]) & 0xC0) != 0x80)) {
14. return false;
15. }
16. }
17. }
18. return true;
19. }

```

Alles anzeigen

== Implementierung und Test ==

Wir können die Funktion beliebig oft auf falsche und richtige String anwenden. Wir erhalten immer einen korrekten UTF-8-String

Quellcode

```

1. $string_correct = utf8_ensure($string_wrong);
2. echo $string_correct;
3. echo utf8_ensure(utf8_ensure(utf8_ensure($string_correct)));
4. function seems_utf8($Str) {
5. for ($i=0; $i<strlen($Str); $i++) {
6. if (ord($Str[$i]) < 0x80) continue; # 0bbbbbbb
7. else if ((ord($Str[$i]) & 0xE0) == 0xC0) $n=1; # 110bbbbbb
8. else if ((ord($Str[$i]) & 0xF0) == 0xE0) $n=2; # 1110bbbb
9. else if ((ord($Str[$i]) & 0xF8) == 0xF0) $n=3; # 11110bbb
10. else if ((ord($Str[$i]) & 0xFC) == 0xF8) $n=4; # 111110bb
11. else if ((ord($Str[$i]) & 0xFE) == 0xFC) $n=5; # 1111110b
12. else return false; // Does not match any model
13. for ($j=0; $j<$n; $j++) {
14. // n bytes matching 10bbbbbb follow ?
15. if ((++$i == strlen($Str)) || ((ord($Str[$i]) & 0xC0) != 0x80)) {
16. return false;
17. }
18. }
19. }
20. return true;
21. }
22. }
23. }
24. return true;
25. }
26. function utf8_ensure($Str) {
27. return seems_utf8($Str)? $Str: utf8_encode($Str);
28. }
29. }

```

Alles anzeigen

== Aktuelle Vorgehensweise ==

Mit den PHP 4 Versionen >= 4.4.3 und den PHP 5 Versionen >= 5.1.3 wurde eine neue Methode eingeführt, die uns das lange Script erleichtert.

Wenn garantiert werden kann, dass diese Versionen installiert sind, kann daher folgende Version verwendet werden:

Quellcode

```

1. function utf8_ensure($Str) {
2. return mb_check_encoding($Str, 'UTF-8') ? $Str : mb_convert_encoding($Str, 'UTF-8', 'auto');
3. }

```