Zeichencodierung

1 ASCII (American Standard Code for Information Interchange)

7 Bit pro Zeichen

- 0 - 31: Steuerzeichen
- 32 - 126: Druckbare Zeichen

```
!"#$%&'()*+,-./0123456789:;<=>?@
ABCDEFGHIJKLMNOPQRSTUVWXYZ
[\]^_`abcdefghijklmnopqrstuvwxyz{|}~
```

- 127: Steuerzeichen

2 ISO 8859

8 Bit pro Zeichen

- 0 - 127: Wie ASCII
- 128 - 255: Je nach Teilnorm

ISO 8859-1 (Latin-1, Westeuropäisch)

z.B. 164: Allgemeines Währungssymbol

ISO 8859-15 (Latin-9, Westeuropäisch)

z.B. 164: Eurosymbol

3 Unicode

Codepunkte und Zeichen

- Jedes Zeichen erhält einen Codepunkt: U+0000, . . . , U+10FFFF
- 17 planes (Ebenen) mit je 216 = 65.536 Zeichen.
  - plane 0: BMP (Basic Multilingual Plane) U+0000 - U+FFFF
    (U+D800 . . . U+DFFF reserviert für Surrogate für Darstellung von höheren Planes in UTF-16)
  - plane 1: U+10000 - U+1FFFF
  - ...
  - plane 16: U+100000 - U+ 10FFFF
- Version 10.0 vom Juni 2017: 136690 Zeichen

Codierung

UTF: Unicode Transformation Format

UTF-32

- Verwendung: ???
- Jedes Unicode-Zeichen wird durch 32 Bit dargestellt.

UTF-16

- Verwendung: Windows, OS X, Java, .NET
- Jedes Unicode-Zeichen wird durch 16 oder 32 Bit dargestellt.
  - Zeichen der BMP: 16 Bit wie Unicode
  - Zeichen außerhalb der BMP: 32 Bit:
    1. von Unicode-Nummer 0x10000 subtrahieren
    2. Ergebnis aufteilen in zwei Blöcke mit je 10 Bit
    3. a) High-Surrogate: vor ersten Block: 110110 ⇒ 0xD800 - 0xDBFF
       b) Low-Surrogate: vor zweiten Block: 110111 ⇒ 0xDC00 - 0xDFFF
    4. High-Surrogat und Low-Surrogat zusammenbauen ⇒ 32 Bit

Beispiele: https://de.wikipedia.org/wiki/UTF-16#Beispiele

UTF-8

- Verwendung: Linux, E-Mail, WWW
- Jedes Unicode-Zeichen wird durch 8, 16, 24 oder 32 Bit dargestellt.

Unicode (hex)	UTF-8 (binär)	Zeichen
00 - 7F	0xxxxxxx	Wie ASCII
80 - 7FF	110xxxxx 10xxxxxx	2 Byte pro Zeichen
800 - FFFF	1110xxxx 10xxxxxx 10xxxxxx	3 Byte pro Zeichen
1000 - 10 FFFF	11110xxx 10xxxxxx 10xxxxxx 10xxxxxx	4 Byte pro Zeichen

4 Sonstiges

BOM (Byte Order Mark)

Bytefolge 0xEF, 0xBB, 0xBF am Dateianfang. Wird nicht von allen Programmen richtig interpretiert.

Kennzeichnung in HTML5

```
<meta charset="utf-8">
```

Zeilenumbrüche

Betriebssystem	Abkürzung	Escape-Squenz	Dezimal
Windows	CR LF	\r \n	13 10
Linux, Mac OS X	LF	\n	10
Mac OS 9	CR LF	\r	13