Unicode

Die Zeichen unserer Tastatur lassen sich binär mit dem 8-Bit-ASCII-Code darstellen.
Allerdings gibt es viele weitere Sprachen, die andere Zeichen und Sonderzeichen verwenden, Zeichensätze mit Piktogrammen und mathematischen Symbolen.

¥ ≠ ♥ ♫ ☼ © Æ ɳ ↑ ∞ µ ֍ ╩ đ ض

Um all diese Zeichen mit einer Codierung darstellen zu können, wurde 1991 der erste Unicode-Standard festgelegt und bis heute kontinuierlich um viele Sprachen und Symbole erweitert. Zuständig dafür ist das Unicode-Konsortium.
Als Unicode-Format hat sich **UTF-8** durchgesetzt.

UTF-16 und UTF-32
sind weitere Unicode-Formate, die mit einer festen Wortlänge von 16 Bit bzw. 32 Bit arbeiten.

Während für eine 8-Bit-ASCII-codierte Nachricht die Wortlänge von 8 Bit pro Zeichen fest vorgegeben ist, kann bei UTF-8 die Wortlänge bei Bedarf vergrößert werden. Für 8-Bit-ASCII und UTF-8 hat das linke Bit in jedem Byte eine besondere Bedeutung:

- In **8-Bit-ASCII** gehört dieses Bit mit zur Zeilenbeschriftung.
  Man erkennt an einer Null, dass es zum 7-Bit-ASCII-Code gehört.
- In **UTF-8** zeigt das Bit an, ob das Zeichen ein oder mehrere Bytes zur binären Darstellung benötigt.
  - Falls linkes Bit = 0:
    die übrigen 7 Bit sind ein 7-Bit-ASCII-Code für ein Zeichen.
    Die ersten 128 von UTF-8 sind also identisch mit denen von 8-Bit-ASCII.
  - Falls linkes Bit = 1:
    dieses Zeichen wird mit mehr als einem Byte dargestellt.
    Das Startbyte beginnt mit 11xxxxxx, die Folgebytes mit 10xxxxxx.
    Man kann an der Anzahl führender Einsen im Startbyte erkennen, wie viele Bytes folgen (maximal noch 3).
    Beginnt das Startbyte mit 110xxxxx, kommt nur noch 1 Folgebyte.
    Es kann aber auch mit 1110xxxx oder 11110xxx beginnen.

8-Bit-ASCII:

Maß	0100 1101	0110 0001	1101 1111

UTF-8:

Maß	0100 1101	0110 0001	1100 0011	1001 1111
☺	1111 0000	1001 1111	1001 1000	1000 1010

Darstellung von UTF-8-Zeichen:
UTF-8-Zeichen werden meist hexadezimal mit vorangestelltem U+ notiert.
Bei der Berechnung werden die roten Ziffern übersprungen.
In Office-Produkten können mit dieser Notation Zeichen eingefügt werden.

M U+004D

a U+0061

ß U+00DF

☺ U+1F60A