Löwenstark Digital Group GmbH
Petzvalstr. 38
38104 Braunschweig
Tel.: 0531 21360 5500
Fax: 0531 21360 5795
kontakt@loewenstark.com

# UTF-8

Stand: 01.08.2022

Beim Thema UFT-8 sind vor allem ITler und Technikspezialisten angesprochen, die entsprechendes Hintergrundwissen mitbringen. UFT-8 steht als Abkürzung für den „Unicode Transformation Format“ auf Basis 8 Bits. Die 8 Bits werden für die Blöcke berechnet, die zur Zeichendarstellung notwendig sind. Pro Zeichen werden für die Darstellung ein bis vier dieser Blöcke benötigt. Die Größe der Dateien hängt von der Länge des Textes und der Art der Zeichen ab. Neben Buchstaben sind auch verschiedene weitere Unicode-Zeichen möglich.

**Zeichenübersicht:**

- Buchstaben des Alphabets
- Zahlen (arabisch), numerische Werte
- Emojis
- weitere Spezial-Symbole wie Einheitendarstellungen, Währungen oder Symbole der Mathematik
- Interpunktion

Über 90 % aller Webseiten arbeiten mit dem UFT-8-Standard. Die Vormachtstellung dieser Darstellung ist vor fast 15 Jahren (2009) erreicht worden.

## Ursprünge von UTF-8

Ken Thompson und Rob Pike gelten als die Väter von UFT-8, welches sie für das sogenannte „Plan-9-Betriebssystem“ entwickelten. Das war schon Anfang der 1990er Jahre. Die Ursprungsbezeichnung bis zur Standardisierung war „FSS-UTF“.

## Das zeichnet UTF-8 aus

Ein dominantes System war lange Zeit ASCII. Die ersten 128 Zeichen von ASCII und UFT-8 sind aus Kompatibilitätsgründen identisch. Damit ist das System sozusagen abwärtskompatibel. Wie erwähnt findet die Kodierung variabel mit unterschiedlicher Länge statt, wobei jeweils 1 bis vier Byte benötigt werden: 1 Byte entspricht genau 8 Bits. Ziel ist jedoch immer die Darstellung mit einem Byte, damit die Dateigröße möglichst begrenzt bleibt. Zeichen-Ökonomie ist hier das Stichwort.

Sein Vorläufer ASCII – „American Standard Code for Information Interchange“ war nur auf das englische Alphabet und diverse Zeichen sowie Satzzeichen und Zahlen ausgelegt. Durch die Ausbreitung des Internets ergab sich jedoch eine zu große Eingrenzung, da die Userzahl stieg, die jeweils etliche Sprachen nutzen und ein einheitlicheres System musste gefunden werden. ASCII wurde im Computerzusammenhang ab 1967 genutzt und hielt sich somit rund 40 Jahre als dominantes System. Es wird teilweise auch heute noch verwendet und gehört zu den Basics in vielen Themenbereichen.

## Funktionsweise UFT-8

Das Bit ist die grundlegende Einheit im Binärsystem. Dieses (bestehend aus Nullen und Einsen) wird von

Löwenstark Digital Group GmbH
Petzvalstr. 38
38104 Braunschweig
Tel.: 0531 21360 5500
Fax: 0531 21360 5795
kontakt@loewenstark.com

Computern verwendet. Damit lassen sich alle Arten von Informationen darstellen. Nach den Bits ist die folgende Größe ein Byte, das aus 8 Bits besteht, beispielsweise: 01110010.

Das UTF-8 oder Unicode Transformation Format stellt eine Erweiterung von ASCII  dar und wandelt sogenannte Codepunkte mit 1 bis 4 Bytes um. Es ist eine simple Form der Codierung. Dank der Codierung sind Zeichen aller Art, also auch Buchstaben fernab des lateinischen Alphabets realisierbar. Die unterschiedlichen Codepoints stellen stellvertretend für Buchstaben und Zeichenkombinationen. Im Unicode wird jedem Codepoint eine Nummer zugeordnet. Hier liegt das Aufgabenfeld der UTF-Codierungen.

Was meint Unicode Transportation Format also? Es geht um Speicher- und Übertragungsformate, mit denen letztlich Unicode-Texte entstehen. Eine unmittelbare Codierungsmöglichkeit für den Unicode ist UTF-32. Durch die Verwendung dieser Variante wird Speicherplatz allerdings manchmal vergeudet. Um dem entgegenzuwirken, entstanden variable Codierungen, die bis heute dominant sind. So werden häufige Zeichen in wenigen Bytes dargestellt und eher seltene Zeichen fordern mehr Platz ein.

- UTF-8
- UTF-16
- UTF-32

## UTF-8-Struktur

- 1 Byte: 128 Zeichen (wie bei den ASCII-Zeichen)
- 2 Bytes: 1920 Zeichen verlangen zwei Bytes für die Kodierung (Arabisch, Griechisch, Hebräisch, Kyrillisch, Latein)
- 3 Bytes: Chinesisch, Japanisch, Koreanisch
- 4 Bytes: Emojis, Einheitenzeichen und mathematische Symbole, historische Schriftzeichen

## Verbreitung von UTF-8 im Netz

E-Mails und Internetseiten werden in allen gängigen Zeichensätzen abgespeichert. In Mails und HTML-Dateien sind Meta-Daten implementiert, die wiederrum Meta-Informationen an die Empfänger übermitteln. Sie werden am Ziel decodiert und ausgelesen. Wenn es doch dazu kommt, dass Codes nicht passgenau sind, müssen Konvertierungen stattfinden. Um das zu vermeiden, wird die Vereinheitlichung immer weiter vorangetrieben.