UTF-8

Was ist UTF-8?

UTF-8 steht für Unicode Transformation Format – 8 Bits. Die „8“ bedeutet, dass zur Darstellung eines Zeichens 8-Bit-Blöcke verwendet werden. Die Anzahl der Blöcke, die zur Darstellung eines Zeichens benötigt werden, variiert zwischen 1 und 4. UTF-8 ist eine kompromisslose Zeichenkodierung, die so kompakt wie ASCII sein kann (wenn die Datei nur aus einem englischen Text besteht), aber auch beliebige Unicode-Zeichen enthalten kann (mit einer gewissen Zunahme der Dateigröße).

Zeichen beziehen sich hier auf Buchstaben in Alphabeten, Zahlen und numerische Werte, Interpunktion, spezielle Symbole (Währungen, mathematische Symbole, Emoji…).

Seit 2009 ist UTF-8 die vorherrschende Kodierung im Internet und macht im März 2020 93,1 % aller Webseiten aus. UTF-8 überholte alle anderen Kodierungen im Jahr 2008 und über 60 % des Internets im Jahr 2012.

Wann entstand UTF-8?

September 1992 wurde die Kodierung von Ken Thompson und Rob Pike bei Arbeiten am Plan-9-Betriebssystem festgelegt. Zuerst wurde es als FSS-UTF bezeichnet. Danach erfolgte durch eine Standardisierung die Namensumbenennung in UTF-8.

Was sind die wichtigsten Eigenschaften von UTF-8?

- **Kompatibilität mit alten Systemen**– UTF-8 muss auf alten Geräten, die nur ASCII verstehen, entzifferbar sein und umgekehrt. Aus diesem Grund sind die ersten 128 englischen Zeichen in UTF-8 die gleichen wie die von ASCII, was es abwärtskompatibel macht.
- **Redundanz**– aus Effizienzgründen verwendet UTF-8 eine Kodierung mit variabler Länge in einem Format von 1 bis 4 Byte (ein Byte entspricht 8 Bits). Wenn ein Zeichen mit nur 1 Byte dargestellt werden kann, werden keine weiteren Bytes hinzugefügt, wodurch die Größe des Dokuments kompakt bleibt.

Was wurde vor UTF-8 verwendet?

Am Anfang gab es ASCII – **A**merican **S**tandard **C**ode for **I**nformation **I**nterchange. ASCII ist ein Zeichencodierungsstandard für die digitale Kommunikation. ASCII umfasst grundlegende Zeichen, Satzzeichen, Zahlen und Buchstaben, die im englischen Alphabet enthalten sind.

Als sich das Internet jedoch immer weiter von den Grundlagen der englischen Sprache entfernte, konnten Milliarden von Nutzern kaum noch lateinische Schriftzeichen verwenden, um auf relevante Inhalte zuzugreifen, und ASCII wurde durch neue Zeichenkodierungen ersetzt.

UTF-8 besitzt die wichtige Eigenschaft, dass ASCII in UTF-8 die gleiche Byte-Kodierung hat wie in der ASCII-Kodierung. Das bedeutet, dass neue Programme mit alten Programmen interagieren können, solange sie nur den ASCII-Zeichensatz verwenden.

Wie funktioniert UTF-8?

Um Informationen zu speichern, verwenden Computer ein binäres System. Im Binärsystem werden alle Daten in einer Folge von 1en und 0en dargestellt. Die grundlegendste Einheit des Binärsystems ist ein Bit, d. h. eine einzelne 1 oder 0. Die nächstgrößere Einheit des Binärsystems, ein Byte, besteht aus 8 Bits. Ein Beispiel für ein Byte ist “01101011”.

In den ersten 128 Zeichen ist UTF-8 mit ASCII deckungsgleich. Es eignet sich mit in der Regel nur einem Byte Speicherbedarf für Zeichen vieler westlicher Sprachen und besonders für die Kodierung englischsprachiger Texte. Diese lässt sich ohne Modifikation, mit nicht UTF-8 fähigen Texteditoren ohne Beeinträchtigung bearbeiten. Das ist einer der Gründe für den Status als De-facto-Standard-Zeichenkodierung des Internet und stellt damit verbundener Dokumenttypen dar.

Wie ist UTF-8 aufgebaut?

UTF-8 oder Unicode Transformation Format ist eine Erweiterung von ASCII. UTF-8 kodiert Codepunkte in ein bis vier Bytes.

Unicode basiert auf sogenannten Codepoints. In einigen Sprachen gibt es Zeichenkombinationen, die einen eigenständigen Buchstaben bilden. Codepoints entsprechen einzelnen Buchstaben, Zeichenkombinationen oder Teilen davon. Jedem Codepoint wird von Unicode eine Nummer zugewiesen, definiert aber nicht, wie diese zu codieren sind. Dafür kommen die UTF-Codierungen zum Einsatz.

UTF steht für Unicode Transportation Format und meint konkrete Speicher- und Übertragungsformate für Unicode-Texte. Die unterschiedlichen UTF-Codierungen spezifizieren, wie die Codepoint-Nummern in Form von Bits und Bytes darzustellen sind. Die direkteste Möglichkeit Unicode zu codieren ist UTF-32, welches jeden Codepoint in 32 Bits sprich 4 Bytes – speichert. Der Nachteil dieser Methode ist die große Speicherplatz-Verschwendung bei Texten, die meist aus dem lateinischen Alphabet bestehen.

Um den Speicherverbrauch zu verbessern, haben sich Formate mit variabler Codierung durchgesetzt. Sie speichern häufig vorkommende Zeichen in wenigen und seltene Zeichen in vielen Bytes.

Während UTF-16 zwei bis vier Bytes verwendet, speichert UTF-8 die häufigsten Zeichen in nur einem Byte. UTF-8 hat sich durchgesetzt, da es ASCII-kompatibel ist. Die ersten 128 Zeichen der UTF-8-Codierung entsprechen genau dem 7-Bit-ASCII-Zeichensatz, sodass reine ASCII-Texte keine Konvertierung erfordern.

Struktur von UTF-8:

- Ein Byte: Die ersten 128 Zeichen (entsprechend den ASCII-Zeichen).
- Zwei Bytes: Die folgenden 1.920 Zeichen benötigen zwei Bytes zur Kodierung (dies umfasst die große Mehrheit der auf lateinischer Schrift basierenden Alphabete, aber auch Hebräisch, Arabisch, Kyrillisch und Griechisch)
- Drei Bytes: Umfasst weitere Zeichen für Sprachen wie Chinesisch, Japanisch, Koreanisch
- Vier Bytes: Enthält historische Schriftzeichen, mathematische Symbole und Emoji.

[i]

Wie sieht UTF-8 aus?

Zeichen	UTF-8 (Hex.)	Unicode
A	41	U+0041
Ä	c3 84	U+00C4
?	f0 9f 98 81	U+1F601
?	e4 b8 ad	U+4E2D

Wie macht sich UTF-8 Im World Wide Web bemerkbar?

SEO-Küche Internet Marketing GmbH & Co. KG
Fraunhoferstr. 6, 83059 Kolbermoor
Telefon 08031 / 2575-100
Telefax 08031 / 2575-101
E-Mail: info@seo-kueche.de

SEO-Küche Internet Marketing GmbH & Co. KG,
HRA 11167 AG Traunstein
pers. Haftende Gesellschafterin:
SEO-Küche Verwaltungs GmbH, Kolbermoor, HRB 22414 AG Traunstein
Geschäftsführer: Christian Brunnenmayer, Patrick Keller, Oliver Lindner
Ust-IdNr.: DE 286 985 708, Steuer Nr. 156/174/08500

HypoVereinsbank
IBAN DE45 700202700015260147
BIC HYVEDEMMXXX

Webseiten und E-Mails lassen sich prinzipiell in allen bekannten Zeichensätzen speichern und übertragen. HTML-Dateien und E-Mails enthalten Metadaten. Diese geben die verwendeten Metainformationen an, sodass die Programme auf Empfängerseite wissen wie die Inhalte zu decodieren sind. Das funktioniert jedoch nicht immer mit allen Programmen. Daher ist eine einheitliche Codierung ratsam, um unnötige Konvertierungen sowie Kompatibilitäts- und Darstellungsprobleme zu vermeiden. Das Internet-Mail Consortium (IMC) und das Word Wide Web Consortium (W3C) empfehlen daher UTF-8 für alle E-Mails und Web-Dateien zu verwenden.

[ii]