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

Unicode

Stand: 15.08.2022

Definition

Unicode ist ein genormter alphanumerischer Zeichensatz, ein sogenanntes Character Encoding Scheme (CES), zur Kodierung von Textzeichen. Unicode umfasst alle weltweit bekannten Textzeichen und enthält auch mathematische, kaufmännische und technische Sonderzeichen. Im Bereich des Online Marketing wird Unicode unter anderem für die HTML-Codierung oder innerhalb von Textverarbeitungsprogrammen angewendet. Darüber hinaus ermöglicht die Zeichensprache die Darstellung von Textzeichen in Form von binären Zahlen, da sie für jedes Zeichen einen Byte-Wert definiert. Die Datenbank für Unicode umfasst derweil um die 230.000 Zeichen und umfasst zudem eine Reserve-Datenbank mit über einer Million weiterer Zeichen. Für die Zuordnung von Zeichen zu Byte-Werten gibt es neben Unicode noch zahlreiche andere, in der Regel unvereinbare Zeichensätze. Der American Standard Code for Information Interchange (ASCII) ist die wohl wichtigste Zeichenkodierung für den digitalen Raum.

Arten von Unicodes

Für die Zuordnung von Zeichen zu Byte-Werten gibt es neben Unicode noch zahlreiche andere, in der Regel unvereinbare Zeichensätze. Der American Standard Code for Information Interchange (ASCII) ist die wohl wichtigste Zeichenkodierung für den digitalen Raum. Innerhalb dieses Zeichensatzes wird jedes Zeichen mit 7-Bits kodiert. Insgesamt können mithilfe des ASCII also 128 Zeichen kodiert werden. Der American Standard Code for Information Interchance beinhaltet im Gegensatz zum Unicode nur die Buchstaben des lateinischen Alphabets und die arabischen Ziffern. Dementsprechend ist die Auszeichnungssprache vor allem für den englischsprachigen Raum nützlich, da der Zeichensatz weder Umlaute noch Akzentzeichen berücksichtigt.

Der Unicode ist in mehrere Ebenen, sogenannte Planes, unterteilt. Dabei wird die erste Ebene, die „Basic Multilingual Plane“ (deutsch: Grundlegende mehrsprachige Ebene) am häufigsten verwendet. Die Zeichensätze auf dieser ersten Ebene werden mithilfe des Universal Character Set 2 (UCS-2) kodiert. Hier werden bereits 16-Bit zur Kodierung jedes Zeichens definiert, sodass insgesamt 65.536 Zeichen verfügbar sind. Statt UCS-2 wird für diese Ebene oft auch der Begriff UFT-16 (UCS Transformation Format 16 Bit) verwendet. Die ersten 265 Zeichen des UFT-16 beinhalten die Schriftzeichen der westeuropäischen Sprachen.

Auf den übrigen Ebenen des Unicodes, die über die erste Ebene hinausgehen, sind selten verwendete, meist historische Schriftzeichen kodiert. Hier finden sich unter anderem alt-ägyptische Hieroglyphen oder seltene chinesische Schriftzeichen. Da 16-Bit für die Kodierung dieser Zeichen nicht mehr ausreichend ist, wird jedes Zeichen mit 32-Bit kodiert, sodass insgesamt 4.294.967.296 verschiedene Zeichen möglich sind. Die höheren Ebenen des Unicodes werden als Universal Character Set 4 (UCS-4) bezeichnet. Die UCS-4-Kodierung ermöglicht die Darstellung jedes beliebigen Unicode-Zeichens unabhängig von der Unicode-Ebene in einem 32-Bit langen Datenwort. UCS-4 wird auch als UTF-32 (UCS Transformation Format 32 Bit) bezeichnet. Bei der Verwendung sollte die der hohe Ressourcenbedarf berücksichtigt werden.

Neben UTF-16 und UTF-32 wird im europäischen Raum vor allem das UCS Transformation Format 8 Bit

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

(UTF-8) angewendet. UTF-8 kann jedes Unicode-Zeichen als Abfolge von Datenwörtern von je 8 Bit Länge ausdrücken und ermöglicht die Umkodierung der Schriftzeichen von 16-Bit auf 8-Bit. UTF-8 stimmt in den ersten 128 Zeichen mit der ASCII-Kodierung überein.

Bedeutung für das Online Marketing

Der Unicode-Standard wird heutzutage schon von führenden internationalen Unternehmen wie Apple, IBM, Microsoft oder Hewlett-Packard verwendet. Auch bei der Programmiersprache Java kommt der Unicode zum Einsatz. Die Kodierung mithilfe des im europäischen Raum gängigen UTF-8 ermöglicht eine hohe Usability der Webseite und eine große mögliche Reichweite, da die verwendeten Zeichen in der Regel weltweit gebräuchlich sind. Darüber hinaus ist UTF-8 im Vergleich zu anderen Unicodes relativ ressourcenschonend. Im Vergleich zum derzeitigen Standard ASCII können mithilfe von UTF-8 auch Umlaute und Akzentzeichen dargestellt werden.