

25 Jahre Unicode

Warum Unicode kein alter Hut ist





- Was ist Unicode?
- Struktur von Unicode
- Spezielle Unicode-Themen
- Skripte (Schriften) und einige Besonderheiten
- Unicode in der öffentlichen Verwaltung
- Zusammenfassung





Was ist Unicode?

- Unicode ist eine Sammlung von über 100.000 Schriftzeichen unterschiedlichster Skripte für wissenschaftliche und kommerzielle Anwendungen
- Unicode definiert eine Vielzahl von Attributen und Algorithmen zum Umgang mit Zeichen und Skripten
- Unicode entspricht im Wesentlichen der ISO 10646 und wird vom Unicode Consortium (www.unicode.org) verwaltet





- Unicode gliedert sich in
 - 17 Ebenen (ab Unicode Version 2.0) mit jeweils 65.536 (0x10000) Zeichen, davon werden derzeit 5 Ebenen verwendet
 - 273 Blöcken mit jeweils bis zu 20.941 Zeichen („Vereinheitlichte CJK-Ideogramme“)
 - mehrere Private Use Areas und spezielle Blöcke für Surrogates
 - 109.449 definierte Zeichen
- Weit mehr als die Hälfte aller Zeichen sind CJK bzw. Han-Zeichen
- Alle Unicode-Zeichen haben einen eindeutigen Namen (z.B. „Latin Capital Letter A“) und einen 4 bis 5-stelligen Hex-Code (z.B. [U+0041](#))
- Neben der Liste der Zeichen (Character Data Base) gibt es eine Vielzahl zusätzlicher Tabellen





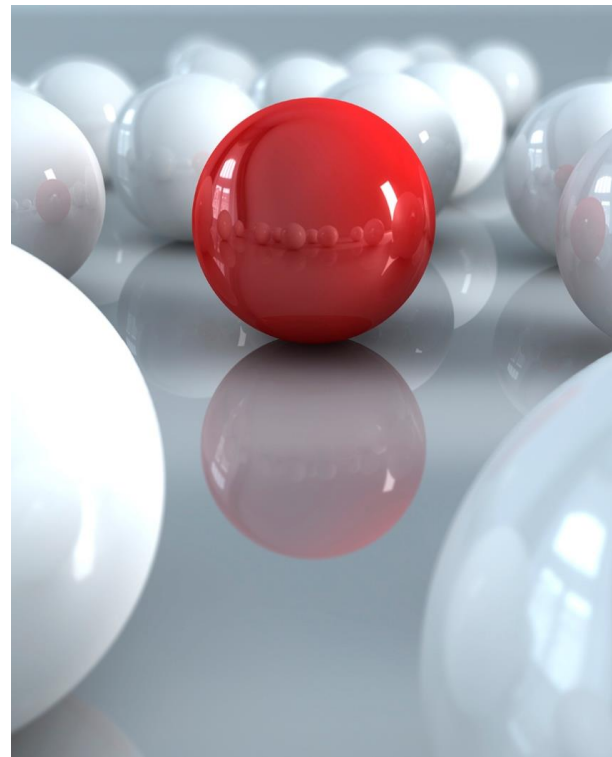
- Die erste Version 1.0 wurde im Oktober 1991 veröffentlicht
- Mit der Version 2.0 im Juli 1996 wurden die Ebenen eingeführt
- Version 2.1 wurde im Mai 1998 eigens vor der Euro-Einführung veröffentlicht
- Mit der Version 6.0 im Oktober 2010 wurden viele Emoji eingeführt
- Seit der Version 7.0 im Juni 2014 wird eine Hauptversion pro Jahr veröffentlicht
- Unicode liegt aktuelle in der **Version 9.0** vor





Attribute der Unicode-Zeichen (Auszug)

- Entsprechender Groß-/Kleinbuchstabe
- Kategorie (z.B. „Großbuchstabe“ oder „Kontroll-Zeichen“)
- Bidirektionalklasse (z.B. „links nach rechts“)
- Verbindungsklasse (z.B. „nicht versetzt“)
- Zerlegung (z.B. ö entsteht aus o+¨)
- Zeichen wird bei bidirektionaler Schreibrichtung gespiegelt
- Schrift (z.B. „lateinisch“)

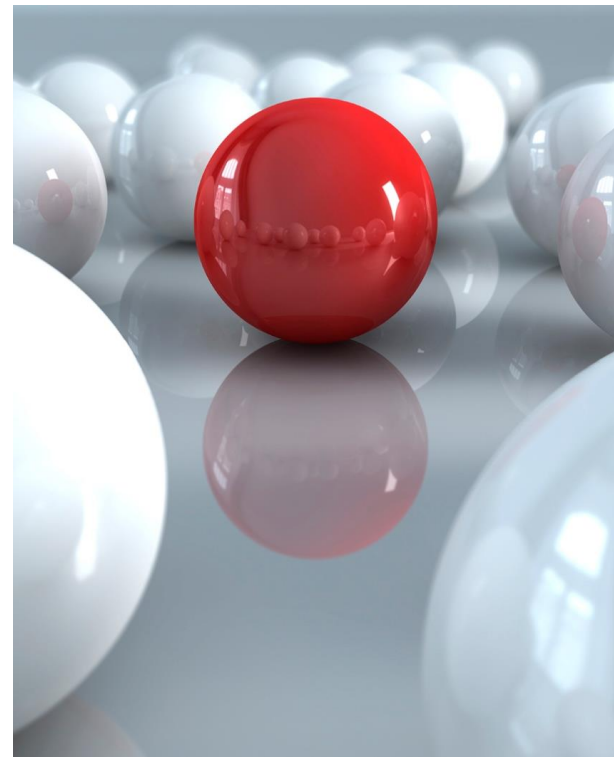




Weitere Attribute anderer Standards

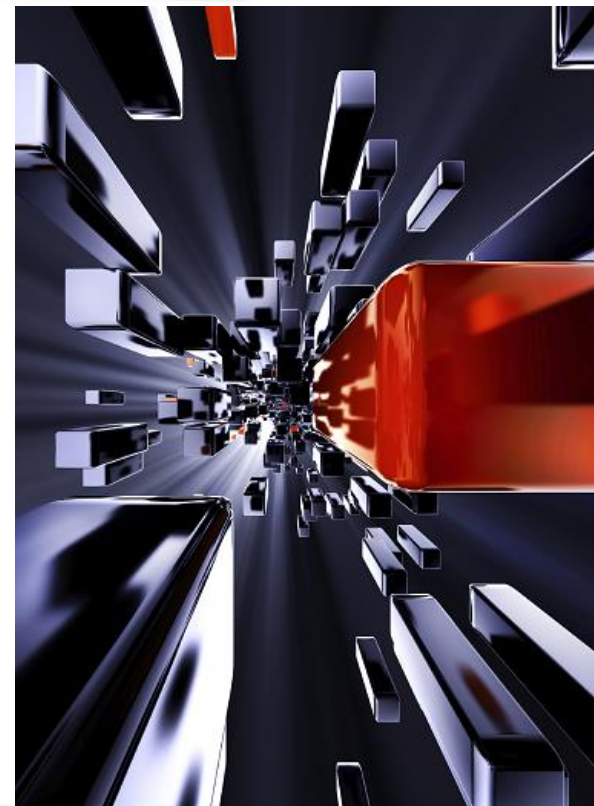
Am Beispiel “Ä”

- AFP's GCGID („LA180000“)
- Postscript Name („Adieresis“)
- HTML-Entity („Auml“)
- Schreibweise in Java/Javascript („\u00C4“)





- Unicode-Texte können über UTF in Byte-Sequenzen codiert werden:
 - UTF-7: heute nahezu bedeutungslos
 - UTF-8: die 21-Bit Unicode Werte werden in 1, 2, 3 oder 4 Byte-Werte gewandelt
 - UTF-16: die 21-Bit Unicode Werte werden in 1 oder 2 16-Bit-Werte im Little Endian (UTF-16LE) oder Big Endian (UTF-16BE) Format gewandelt (Surrogates)
 - UTF-32: Pure Platzverschwendung
- Die meisten Systeme (Java, Javascript, Word) arbeiten intern im UTF16-Format



- Unicode selbst kennt keine Codepages, das Unicode Consortium veröffentlicht aber eine Vielzahl von Codepages mit den entsprechenden Unicode-Codes
- XML, HTML, HTTP und viele andere Systeme benutzen zur Codierung entweder Codepage-Namen (durch IANA definiert) oder UTF-Methoden, die dann als „Encoding“ bezeichnet werden, z.B. „ISO-8859-1“, „IBM500“ oder „UTF-8“
- Der erste Unicode-Block mit 128 Zeichen entspricht dem ASCII-Zeichensatz
- Die ersten beiden Unicode-Blöcke mit zusammen 256 Zeichen entsprechen der ISO Latin 1 Codepage





- True Type (TTF) und Open Type Fonts (OTF) unterstützen Unicode (zumindest bis Version 2)
- Es gibt keine Unicode-Fonts oder besser, Fonts, die alle Unicode-Zeichen beinhalten
- Streng genommen muss jedes Unternehmen eine Liste der unterstützten Unicode-Zeichen erstellen
- Es gibt eine Vielzahl sehr umfassender freier Fonts, z.B:
 - Linux Libertine
 - Bitstream Vera
 - Noto (Google und Adobe)
 - Arial Unicode MS (Bestandteil von Office)





- Die ISO 15924 definiert über 180 Skripte mit Nummern und Abkürzungen
- Beispiele:
 - Zyyy (998): **Allgemein**, 7129 Zeichen in Unicode
 - Latn (215): **Lateinisch**, 1338 Zeichen in Unicode
 - Cyrl (431): **Kyrillisch**, 431 Zeichen in Unicode
 - Brai (570): **Braille**, 256 Zeichen in Unicode





- Diakritika sind Zeichen ohne eigene Breite, die mit anderen Zeichen verbunden werden. Beispiel: Ö (U+00D6) wird zusammengesetzt aus O (U+004F) und ¨ (U+0308).
- Beispiele:
 - ´ Akut (U+0301), z.B. é
 - ` Gravis (U+0300), z.B. è
 - ^ Zirkumflex (U+0302), z.B. ê
 - ˘ Breve (U+0306), z.B. ġ
 - ¨ Trema (U+0308), z.B. ö
 - , Cedille (U+0327), z.B. ç



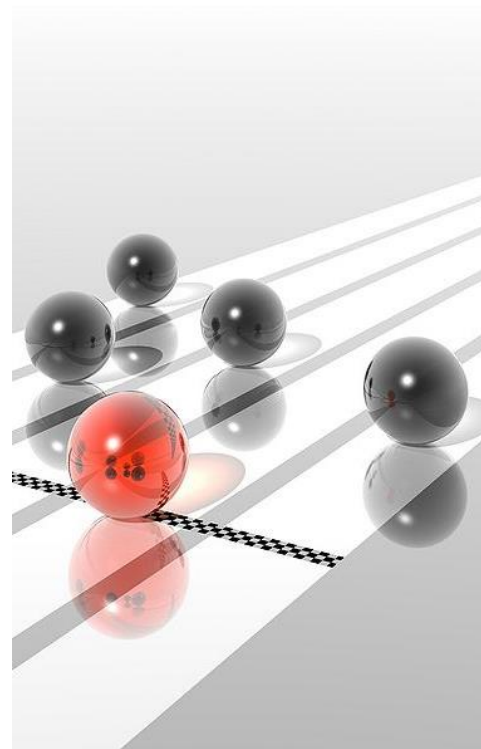


- Die Koordinierungsstelle für IT-Standards in Bremen veröffentlicht eine Liste der „Lateinischen Zeichen in Unicode“ die der IT-Planungsrat des Bundesministeriums des Inneren als verbindlich beschlossen hat
- Diese Liste (http://xoev.de/latinchars/1_1/latinchars.pdf) umfasst 36 Sonderzeichen und 454 Varianten lateinischer Buchstaben
- Die lateinische Schreibweise ist Ergebnis einer Transliteration (buchstabengetreue Übertragung aus einer Schrift in eine andere)
- Sie soll unter anderem das Auffinden von Personennamen im Melderegister vereinfachen (z.B. Möller vs. Møller)





- An Unicode führt kein Weg vorbei
- Nach 25 Jahren Unicode könnte man langsam anfangen, sich damit zu beschäftigen
- Codepages sind ein Relikt aus den 90er Jahre
- Fonts sind und bleiben eine kleine Herausforderung
- Spruch aus der Wikipedia: Sei mutig!





Weitere Informationen:

<https://www.compart.com/de/unicode>

Harald Grumser

Compart AG

Otto-Lilienthal-Str. 38

71034 Böblingen

Tel.: +49 7031 6205-0

E-Mail: Harald.Grumser@compart.com

