Technische Informatik I
Zeichenkodierung

Thorsten Thormählen
31. Oktober 2023
Teil 2, Kapitel 3

Dies ist die Druck-Ansicht.

Aktiviere Präsentationsansicht

Steuerungstasten

→ nächste Folie (auch Enter oder Spacebar).
← vorherige Folie
d schaltet das Zeichnen auf Folien ein/aus
p wechselt zwischen Druck- und Präsentationsansicht
CTRL + vergrößert die Folien
CTRL - verkleinert die Folien
CTRL 0 setzt die Größenänderung zurück

Inhalt

ASCII-Code

ISO 8859

Unicode

Zeichenkodierung

Um Text auf einem Computer darzustellen, muss jeder Buchstabe binär kodiert werden

Je nachdem wie viele Bits pro Zeichen verwendet werden, können unterschiedlich viele verschiedene Zeichen abgelegt werden

Beispiel:

7 Bits: $2^7 = 128$ verschiedene Zeichen

8 Bits: $2^8 = 256$ verschiedene Zeichen

16 Bits: $2^{16} = 65536$ verschiedene Zeichen

ASCII-Code

Der ASCII-Code (American Standard Code for Information Interchange) ist eine 7-Bit-Zeichenkodierung, die 1963 von der American Standards Association (ASA) beschlossen wurde

Ein Zeichen wird jedoch immer als 1 Byte (=8 Bits) abgelegt, d.h. das höchstwertige (8.) Bit ist immer Null

Insgesamt gibt es 128 Zeichen, davon 95 druckbare und 33 Steuerzeichen

ASCII-Code

In folgender Tabelle sind alle 128 ASCII-Zeichen angegeben

Wird der Mauszeiger über ein Zeichen gehalten, wird eine kurze Beschreibung angezeigt

Hex Code	...0	...1	...2	...3	...4	...5	...6	...7	...8	...9	...A	...B	...C	...D	...E	...F
0...	*NUL*	*SOH*	*STX*	*ETX*	*EOT*	*ENQ*	*ACK*	*BEL*	*BS*	*HT*	*LF*	*VT*	*FF*	*CR*	*SO*	*SI*
1...	*DLE*	*DC1*	*DC2*	*DC3*	*DC4*	*NAK*	*SYN*	*ETB*	*CAN*	*EM*	*SUB*	*ESC*	*FS*	*GS*	*RS*	*US*
2...	SP	!	"	#	$	%	&	'	(	)	*	+	,	-	.	/
3...	0	1	2	3	4	5	6	7	8	9	:	;	<	=	>	?
4...	@	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
5...	P	Q	R	S	T	U	V	W	X	Y	Z	[	\	]	^	_
6...	`	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o
7...	p	q	r	s	t	u	v	w	x	y	z	{	|	}	~	DEL

Beispiel: Das Zeichen "A" hat den Hexadezimalwert $(41)_{16}$

$(41)_{16} = (01000001)_2 = (65)_{10}$

[Quelle: Wikipedia]

ASCII-Code

Die Steuerzeichen stammen aus einer Zeit, in der der ASCII-Code zur Steuerung von Fernschreibern (elektrisch angesteuerte Schreibmaschinen) verwendet wurde

Heutzutage habe viele dieser Steuerzeichen ihre Bedeutung verloren

Wichtig ist eigentlich nur noch das Steuerzeichen für eine neue Zeile: "LF" (Line Feed, ASCII $(0A)_{16}$)

Beim Betriebsystem Windows muss dem "Line Feed" Zeichen allerdings noch ein "Carriage Return" vorangestellt werden: "CR LF" (=ASCII $(0D)_{16}$ $(0A)_{16}$)

ISO 8859

Bei der ASCII-Codierung werden nur 7 der 8 Bits eines Bytes genutzt

Der restliche Zahlenbereich (128 bis 255) kann also für weitere Zeichen verwendet werden

Die *International Organization for Standardization* definiert in ISO 8859 insgesamt 15 ASCII-Erweiterungen

ISO 8859-1 enthält z.B. die für uns in Deutschland wichtigen Buchstaben: "ä", "ö", "ü", "ß"

ISO 8859-1	Westeuropäisch (Latin-1)
ISO 8859-2	Mitteleuropäisch (Latin-2)
ISO 8859-3	Südeuropäisch (Latin-3)
ISO 8859-4	Nordeuropäisch (Latin-4)
ISO 8859-5	Kyrillisch
ISO 8859-6	Arabisch
ISO 8859-7	Griechisch
ISO 8859-8	Hebräisch
ISO 8859-9	Türkisch (Latin-5)
ISO 8859-10	Nordisch (Latin-6)
ISO 8859-11	Thai
ISO 8859-12	verworfen
ISO 8859-13	Baltisch (Latin-7)
ISO 8859-14	Keltisch (Latin-8)
ISO 8859-15	Westeuropäisch (Latin-9)
ISO 8859-16	Südosteuropäisch (Latin-10)

Unicode

Bei der Verwendung von ISO 8859 zum Austausch von Texten kommt es immer wieder zu fehlerhaften Darstellungen von Zeichen. Dies passiert leicht, wenn Sender und Empfänger nicht die gleiche ISO 8859-x Norm zur Dekodierung verwenden

Ausserdem sind in ISO 8859 längst nicht alle Schriftzeichen aus den unterschiedlichsten Kulturkreisen erfasst

Die Bestrebung des *Unicode* ist es, eine einzige universelle Kodierung zu definieren, die alle relevanten Zeichen enthält

Der Unicode wurde von der ISO als ISO-10646 standardisiert

Unicode

Der Unicode besteht aus 17 Ebenen (darstellbar mit 5 Bits).

Jede Ebene hat 16 Bits und kann damit theoretisch

2^{16}=65536 Zeichen kodieren

Insgesamt kann ein Unicode also 5+16=21 Bits benötigen

Die meisten aktuell verwendeten Zeichen sind in Ebene 0, der Basic Multilingual Plane (BMP), zu finden

Ein Unicode Zeichen wird üblicherweise als ein "U+" und einer Hexadzimalzahl mit mindestens 4 Stellen angegeben

Beispiele:
U+00E4 für das "ä"

U+1300C für die ägyptische Hieroglyphe 

[Bildquelle: Unicodeblock Ägyptische Hieroglyphen]

Unicode

Die Kodierung aller möglichen Schriftzeichen ist ein andauernder Prozess, d.h. die Anzahl der Zeichen wächst ständig

Ein Problem bei der Darstellung ist, dass die meisten Schriftarten nur eine kleine Untermenge der im Unicode definierten Zeichen bereit halten

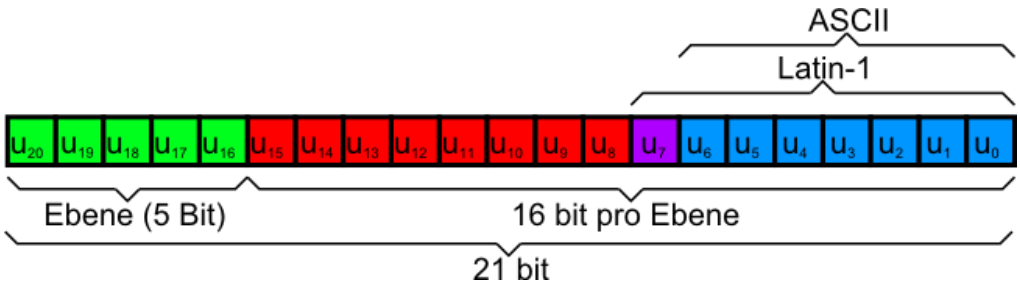
Ist ein Zeichen in einer Schrift nicht vorhanden, wird oftmals einfach ein Zeichen aus einer anderen Schriftart eingefügt

Die Webseite http://www.decodeunicode.org/ hat es sich zur Aufgabe gemacht, alle aktuell im Unicode kodierten Zeichen darzustellen

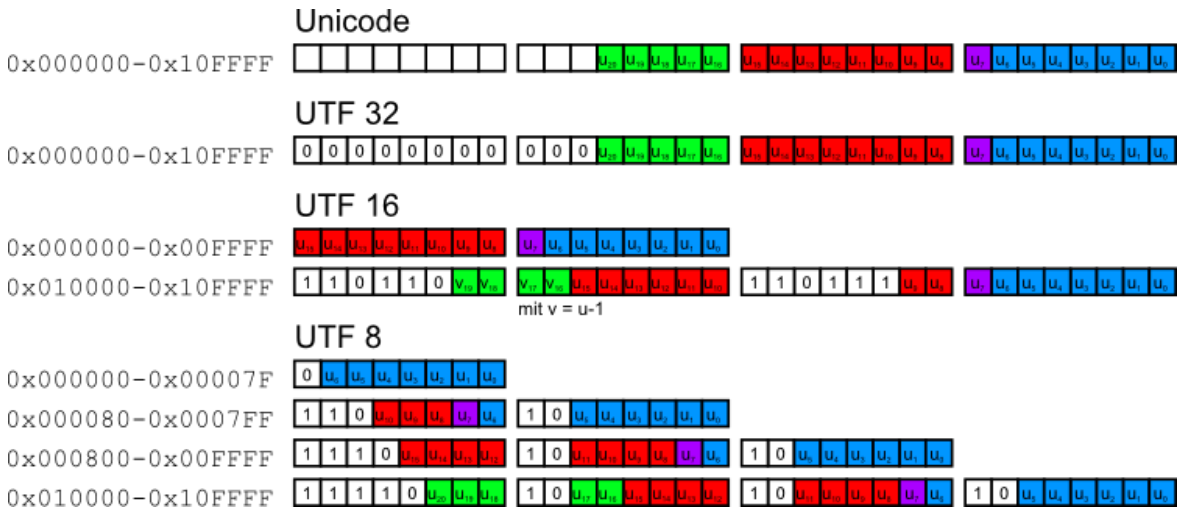
Unicode

Beim Entwurf des Unicode wurde auf Kontinuität wert gelegt

Aus den 21 Bits des Unicodes entsprechen die ersten 7 Bits dem ASCII-Code und die ersten 8 Bits der ASCII-Erweiterungen ISO 8859-1 (Latin 1)



UTF

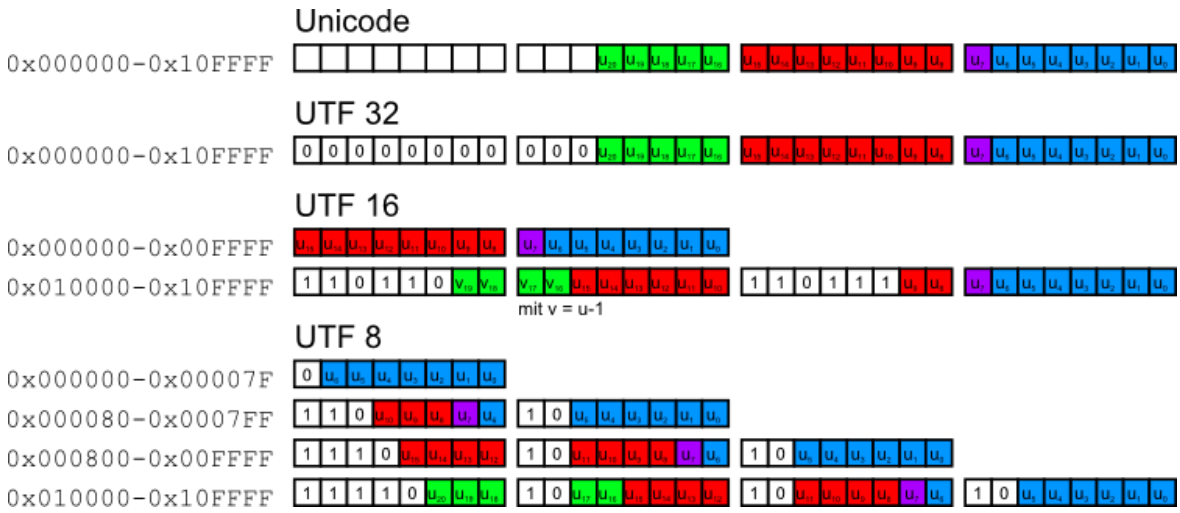


Zur Kodierung von Unicode-Zeichen wird meistens das UTF "Universal Transformation Format" verwendet

UTF-32 kodiert jedes Unicode-Zeichen mit 32 Bits, indem es die 21 Unicode Bits mit Nullen auffüllt

UTF-16 kodiert alle Bits der Basic Multilingual Plane (BMP) mit 16 Bits, nur für die anderen Ebenen werden 32 Bits benötigt

UTF-8

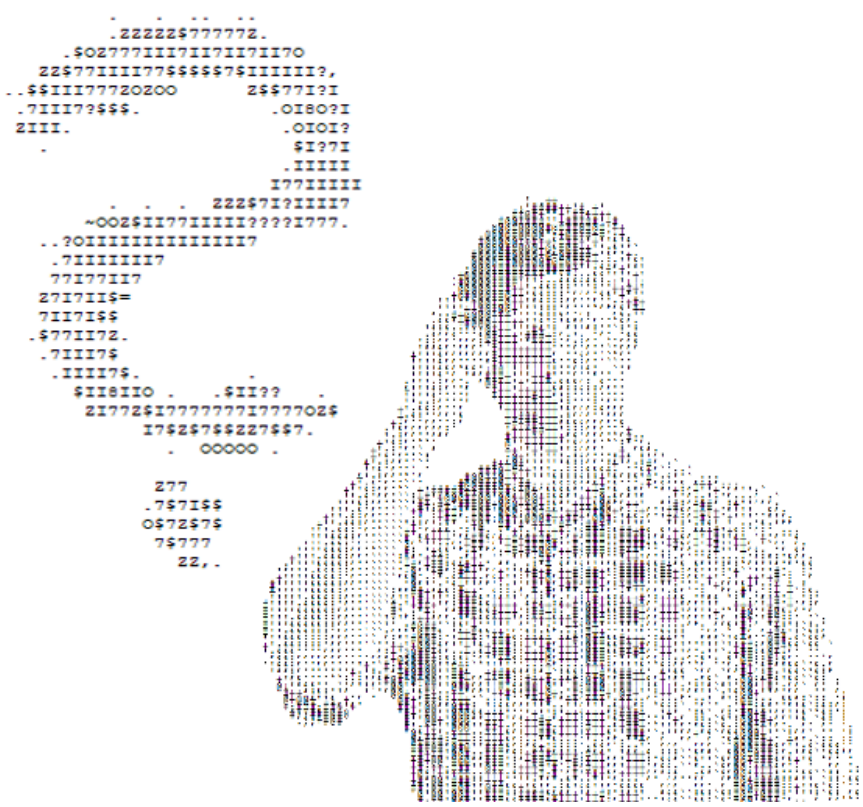


UTF-8 kodiert die ersten 7 Unicode Bits (entspricht ASCII) mit 8 Bits, die ersten 11 Unicode Bits mit 16 Bits, usw.

Ein UTF-8 kodierter Text, der nur ASCII Zeichen enthält, ist demnach vollständig mit ASCII kompatibel

UTF-8 ist heutzutage (besonders im Internet) weit verbreitet (Quasi-Standard der Zeichenkodierung)

Gibt es Fragen?



Gibt es Fragen?

Anregungen oder Verbesserungsvorschläge können auch gerne per E-mail an mich gesendet werden: Kontakt

Weitere Vorlesungsfolien