Zeichenkodierung

Benjamin Tröster

Hochschule für Technik und Wirtschaft Berlin

22. Dezember 2021

Fahrplan

Zeichenkodierung

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

ASCII TABLE

Decimal	Hex	Char	Decimal	Hex	Char	Decimal	Hex	Char	Decimal	Hex	Char
0	0	[NULL]	32	20	[SPACE]	64	40	@	96	60	`
1	1	[START OF HEADING]	33	21	!	65	41	A	97	61	a
2	2	[START OF TEXT]	34	22	"	66	42	B	98	62	b
3	3	[END OF TEXT]	35	23	#	67	43	C	99	63	c
4	4	[END OF TRANSMISSION]	36	24	$	68	44	D	100	64	d
5	5	[ENQUIRY]	37	25	%	69	45	E	101	65	e
6	6	[ACKNOWLEDGE]	38	26	&	70	46	F	102	66	f
7	7	[BELL]	39	27	'	71	47	G	103	67	g
8	8	[BACKSPACE]	40	28	(	72	48	H	104	68	h
9	9	[HORIZONTAL TAB]	41	29	)	73	49	I	105	69	i
10	A	[LINE FEED]	42	2A	*	74	4A	J	106	6A	j
11	B	[VERTICAL TAB]	43	2B	+	75	4B	K	107	6B	k
12	C	[FORM FEED]	44	2C	,	76	4C	L	108	6C	l
13	D	[CARRIAGE RETURN]	45	2D	-	77	4D	M	109	6D	m
14	E	[SHIFT OUT]	46	2E	.	78	4E	N	110	6E	n
15	F	[SHIFT IN]	47	2F	/	79	4F	O	111	6F	o
16	10	[DATA LINK ESCAPE]	48	30	0	80	50	P	112	70	p
17	11	[DEVICE CONTROL 1]	49	31	1	81	51	Q	113	71	q
18	12	[DEVICE CONTROL 2]	50	32	2	82	52	R	114	72	r
19	13	[DEVICE CONTROL 3]	51	33	3	83	53	S	115	73	s
20	14	[DEVICE CONTROL 4]	52	34	4	84	54	T	116	74	t
21	15	[NEGATIVE ACKNOWLEDGE]	53	35	5	85	55	U	117	75	u
22	16	[SYNCHRONOUS IDLE]	54	36	6	86	56	V	118	76	v
23	17	[ENG OF TRANS. BLOCK]	55	37	7	87	57	W	119	77	w
24	18	[CANCEL]	56	38	8	88	58	X	120	78	x
25	19	[END OF MEDIUM]	57	39	9	89	59	Y	121	79	y
26	1A	[SUBSTITUTE]	58	3A	:	90	5A	Z	122	7A	z
27	1B	[ESCAPE]	59	3B	;	91	5B	[	123	7B	{
28	1C	[FILE SEPARATOR]	60	3C	<	92	5C	\	124	7C	|
29	1D	[GROUP SEPARATOR]	61	3D	=	93	5D	]	125	7D	}
30	1E	[RECORD SEPARATOR]	62	3E	>	94	5E	^	126	7E	~
31	1F	[UNIT SEPARATOR]	63	3F	?	95	5F	_	127	7F	[DEL]

- Beispiel: Das Zeichen „A“ hat den Hexadezimalwert 41_{16}

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

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 Betriebssystem 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, ü, ö, 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

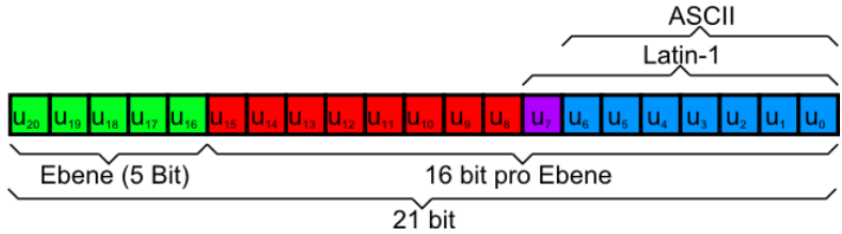
- Der Unicode besteht aus 17 Ebenen (darstellbar mit 5 Bits)
- Jede Ebene hat 16 Bits und kann damit theoretisch $2^{16} = 65536$ Zeichen kodieren
- Insgesamt kann ein Unicode also 5+16=21 Bits benötigen
- Die meisten aktuell verwendeten Zeichen sind in Ebene 0, der Basic Multilingual Plane (BMP), zu finden
- Ein Unicode Zeichen wird üblicherweise als ein „U+“ und einer Hexadezimalzahl mit mindestens 4 Stellen angegeben
- Beispiele:
  - $U + 00E4$ für das ä
  - $U + 00A9$ für die Copyright Symbol ©

Unicode Aufbau

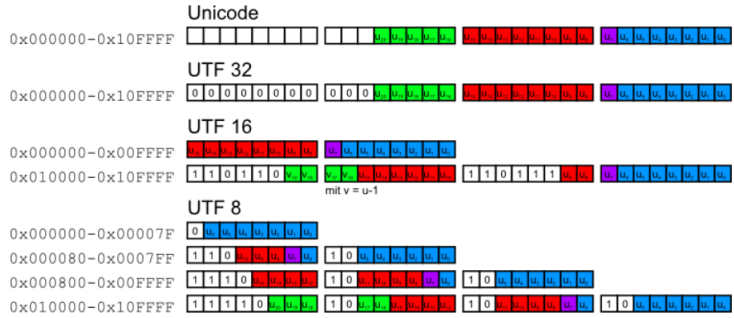
- Die Kodierung aller möglichen Schriftzeichen ist ein andauernder Prozess, d.h. die Anzahl der Zeichen wächst ständig
- Ein Problem bei der Darstellung ist, dass die meisten Schriftarten nur eine kleine Untermenge der im Unicode definierten Zeichen bereit halten
- Ist ein Zeichen in einer Schrift nicht 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)

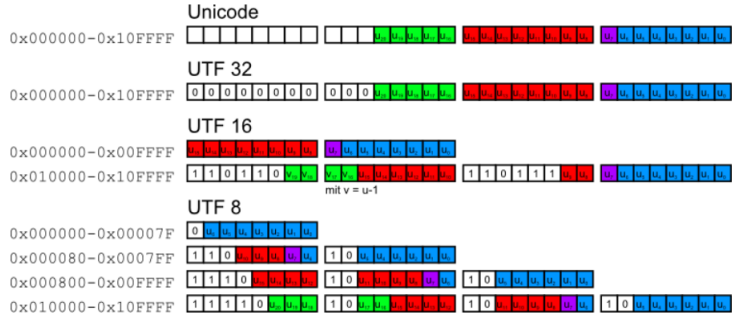


UTF



- 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)

Quellen I