Unicode

Ein kompakter Überblick des Standards

Geschichte, I

- Computer können nur Zahlen
  - Textzeichen repräsentiert durch Zahlen
- ASCII
  - alphanumerische + Steuer- & Symbolzeichen
  - 7-bit Muster

USASCII code chart

Bits $b_7 b_6 b_5$ →	0 0 0	0 0 1	0 1 0	0 1 1	1 0 0	1 0 1	1 1 0	1 1 1
b_4 ↓	b_3 ↓	b_2 ↓	b_1 ↓	Row ↓ / Column → 0	1	2	3	4	5	6	7
0	0	0	0	0	NUL	DLE	SP	0	@	P	`	p
0	0	0	1	1	SOH	DC1	!	1	A	Q	a	q
0	0	1	0	2	STX	DC2	"	2	B	R	b	r
0	0	1	1	3	ETX	DC3	#	3	C	S	c	s
0	1	0	0	4	EOT	DC4	$	4	D	T	d	t
0	1	0	1	5	ENQ	NAK	%	5	E	U	e	u
0	1	1	0	6	ACK	SYN	&	6	F	V	f	v
0	1	1	1	7	BEL	ETB	'	7	G	W	g	w
1	0	0	0	8	BS	CAN	(	8	H	X	h	x
1	0	0	1	9	HT	EM	)	9	I	Y	i	y
1	0	1	0	10	LF	SUB	*	:	J	Z	j	z
1	0	1	1	11	VT	ESC	+	;	K	[	k	{
1	1	0	0	12	FF	FS	,	<	L	\	l	|
1	1	0	1	13	CR	GS	-	=	M	]	m	}
1	1	1	0	14	SO	RS	.	>	N	^	n	~
1	1	1	1	15	SI	US	/	?	O	_	o	DEL

http://en.wikipedia.org/wiki/File:ASCII_Code_Chart-Quick_ref_card.png

Geschichte, 2

- Text Encodings/Code Pages (8 bit)
  - Kollisionen
    - z.B. Westeuropäisch und Griechisch
  - Kompatibilitätsprobleme

ISO/IEC 8859-1

	_0	_1	_2	_3	_4	_5	_6	_7	_8	_9	_A	_B	_C	_D	_E	_F
0_																
1_																
2_	SP 32	! 33	" 34	# 35	$ 36	% 37	& 38	' 39	(40	) 41	* 42	+ 43	, 44	- 45	. 46	/ 47
3_	0 48	1 49	2 50	3 51	4 52	5 53	6 54	7 55	8 56	9 57	: 58	; 59	< 60	= 61	> 62	? 63
4_	@ 64	A 65	B 66	C 67	D 68	E 69	F 70	G 71	H 72	I 73	J 74	K 75	L 76	M 77	N 78	O 79
5_	P 80	Q 81	R 82	S 83	T 84	U 85	V 86	W 87	X 88	Y 89	Z 90	[91	\ 92	] 93	^ 94	_ 95
6_	` 96	a 97	b 98	c 99	d 100	e 101	f 102	g 103	h 104	i 105	j 106	k 107	l 108	m 109	n 110	o 111
7_	p 112	q 113	r 114	s 115	t 116	u 117	v 118	w 119	x 120	y 121	z 122	{ 123	| 124	} 125	~ 126	
8_																
9_																
A_	NBSP 160	¡ 161	¢ 162	£ 163	¤ 164	¥ 165	¦ 166	§ 167	¨ 168	© 169	ª 170	« 171	¬ 172	SHY 173	® 174	¯ 175
B_	° 176	± 177	² 178	³ 179	´ 180	µ 181	¶ 182	· 183	¸ 184	¹ 185	º 186	» 187	¼ 188	½ 189	¾ 190	¿ 191
C_	À 192	Á 193	Â 194	Ã 195	Ä 196	Å 197	Æ 198	Ç 199	È 200	É 201	Ê 202	Ë 203	Ì 204	Í 205	Î 206	Ï 207
D_	Ð 208	Ñ 209	Ò 210	Ó 211	Ô 212	Õ 213	Ö 214	× 215	Ø 216	Ù 217	Ú 218	Û 219	Ü 220	Ý 221	Þ 222	ß 223
E_	à 224	á 225	â 226	ã 227	ä 228	å 229	æ 230	ç 231	è 232	é 233	ê 234	ë 235	ì 236	í 237	î 238	ï 239
F_	ð 240	ñ 241	ò 242	ó 243	ô 244	õ 245	ö 246	÷ 247	ø 248	ù 249	ú 250	û 251	ü 252	ý 253	þ 254	ÿ 255
	_0	_1	_2	_3	_4	_5	_6	_7	_8	_9	_A	_B	_C	_D	_E	_F

Mac Roman

	_0	_1	_2	_3	_4	_5	_6	_7	_8	_9	_A	_B	_C	_D	_E	_F
0_	NUL 0	SOH 1	STX 2	ETX 3	EOT 4	ENQ 5	ACK 6	BEL 7	BS 8	HT 9	LF 10	VT 11	FF 12	CR 13	SO 14	SI 15
1_	DLE 16	DC1⌘ 17	DC2⌘ 18	DC3⌘ 19	DC4⌘ 20	NAK 21	SYN 22	ETB 23	CAN 24	EM 25	SUB 26	ESC 27	FS 28	GS 29	RS 30	US 31
2_	SP 32	! 33	" 34	# 35	$ 36	% 37	& 38	' 39	(40	) 41	* 42	+ 43	, 44	- 45	. 46	/ 47
3_	0 48	1 49	2 50	3 51	4 52	5 53	6 54	7 55	8 56	9 57	: 58	; 59	< 60	= 61	> 62	? 63
4_	@ 64	A 65	B 66	C 67	D 68	E 69	F 70	G 71	H 72	I 73	J 74	K 75	L 76	M 77	N 78	O 79
5_	P 80	Q 81	R 82	S 83	T 84	U 85	V 86	W 87	X 88	Y 89	Z 90	[91	\ 92	] 93	^ 94	_ 95
6_	` 96	a 97	b 98	c 99	d 100	e 101	f 102	g 103	h 104	i 105	j 106	k 107	l 108	m 109	n 110	o 111
7_	p 112	q 113	r 114	s 115	t 116	u 117	v 118	w 119	x 120	y 121	z 122	{ 123	| 124	} 125	~ 126	DEL 127
8_	Ä 128	Å 129	Ç 130	É 131	Ñ 132	Ö 133	Ü 134	á 135	à 136	â 137	ä 138	ã 139	å 140	ç 141	é 142	è 143
9_	ê 144	ë 145	í 146	ì 147	î 148	ï 149	ñ 150	ó 151	ò 152	ô 153	ö 154	õ 155	ú 156	ù 157	û 158	ü 159
A_	† 160	° 161	¢ 162	£ 163	§ 164	• 165	¶ 166	ß 167	® 168	© 169	™ 170	´ 171	¨ 172	≠ 173	Æ 174	Ø 175
B_	∞ 176	± 177	≤ 178	≥ 179	¥ 180	μ 181	∂ 182	∑ 183	∏ 184	π 185	∫ 186	ª 187	º 188	Ω 189	æ 190	ø 191
C_	¿ 192	¡ 193	¬ 194	√ 195	ƒ 196	≈ 197	Δ 198	« 199	» 200	… 201	NBSP 202	À 203	Ã 204	Õ 205	Œ 206	œ 207
D_	– 208	— 209	“ 210	” 211	‘ 212	’ 213	÷ 214	◊ 215	ÿ 216	Ÿ 217	⁄ 218	€¤ 219	‹ 220	› 221	fi 222	fl 223
E_	‡ 224	· 225	‚ 226	„ 227	‰ 228	Â 229	Ê 230	Á 231	Ë 232	È 233	Í 234	Î 235	Ï 236	Ì 237	Ó 238	Ô 239
F_	 240	Ò 241	Ú 242	Û 243	Ù 244	ı 245	ˆ 246	˜ 247	¯ 248	˘ 249	˙ 250	˚ 251	¸ 252	˝ 253	˛ 254	ˇ 255
	_0	_1	_2	_3	_4	_5	_6	_7	_8	_9	_A	_B	_C	_D	_E	_F

Geschichte, 3

- Unicode
  - standardisiert ab 1991
- Untermenge UCS

Universal Character Set (UCS), 1

- UCS: einfache Zuordnung von
  - Symbolen
  - Namen (in Großbuchstaben)
  - Zahlen (Unicode Code Points, U+HEX)
- „e“ U+0065 (LATIN SMALL LETTER E)

Universal Character Set (UCS), 2

- ursprünglich 16-bit (65,536)
  - UCS-2
- U+0000 bis U+FFFF

Erweiterung auf 21 bit, 1

- Zeichenvorrat zu klein
  - historische und selten verwendete Zeichen
- UTF-16: U+0000 bis U+10FFFF
- 1,114,112 Zeichen verteilt auf
  - 17 Ebenen (Planes) à 65,536 Zeichen

Erweiterung auf 21 bit, 2

- ursprüngliche Zeichen in Plane 0
  - Basic Multilingual Plane (BMP)
- Emoji in Plane 1
  - 😛

Private Use Area (PUA)

- Nutzungsbereich für „eigene“ Zeichen
- Ursprünglich U+E000 bis U+F8FF
- Zusätzlich: Plane 15 & 16
- nicht im Datenaustausch verwenden
  - nur „intern“!

Unicode

- Zusätzlich zu UCS:
  - Sortierungsregeln
  - Algorithmensammlung
  - Metadaten
    - „character properties“

Rückwärtskompatibilität

- die ersten 256 Code Points entsprechen ISO Latin 1 (ISO-8859-1)
- Combining Character Sequences

Combining Character Sequences, I

- mehrere Wege führen ~~nach Rom~~ zu einigen der Zeichen
- Als einzelnes Symbol …
  - „é“ U+00E9 (LATIN SMALL LETTER E WITH ACUTE)

Combining Character Sequences, 2

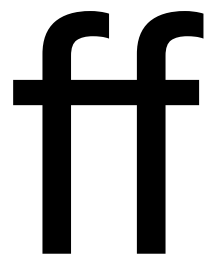
- … oder zusammengesetzt aus mehreren:
  - „e“ U+0065 (LATIN SMALL LETTER E) +
  - „◌́“ U+0301 (COMBINING ACUTE ACCENT)

Equal vs. equivalent

- die beiden „é“ sind nicht „gleich“ sondern „canonically equivalent“
- manche Zeichen sind visuell identisch, aber tragen unterschiedlicher Bedeutungen:
  - Å (LATIN CAPITAL LETTER A WITH RING ABOVE, U+00C5)
  - Å (ANGSTROM SIGN, U+212B)

Beispiel: Ligaturen

- zu einem Zeichen zusammengezogene Buchstaben
- (LATIN SMALL LIGATURE FF, U+FB00)
- (LATIN SMALL LETTER F * 2, U+0066)
- compatibility equivalence
  - (aber nicht „canonically equivalent“)



Unicode Code Points

- … sind nur einfache Zahlen
- wir brauchen deshalb
  - Unicode Transformation Format (UTF)
  - algorithmisch festgelegte Kodierungen eines jeden Unicode Code Point als Bytesequenz

Unicode Transformation Formats (UTF), 1

- verbreitet:
  - UTF-8
  - UTF-16
  - UTF-32
  - plus maschinenabhängige Varianten (endianness, BOM)

Unicode Transformation Formats (UTF), 2

- UTF-8
  - Web, Go, Linux, Unix
- UTF-16
  - APIs von Java, JavaScript, OS X, Windows
- UTF-32
  - Linux, Unix

Weitere wichtige Begriffe

Code Units

- „Datenwörter“ fester Größe
- manchmal mehrere Code Units für einen Code Point nötig
- Besonders bei UTF-8 …
- … aber auch bei UTF-16!
- Ausnahme: UTF-32

Grapheme Clusters

- „user-perceived characters“
  - ≥ 1 Code Point
    - auch in UTF-32 (combining characters)!
- Segmentierungalgorithmus

Weggelassen

- Normalization Forms
- Glyphenvarianten (z.B. Arabisch)

Quellen, I

- NSString and Unicode - Strings - objc.io issue #9
- Unicode - Wikipedia, the free encyclopedia
- Private Use Areas - Wikipedia, the free encyclopedia
- Unicode equivalence - Wikipedia, the free encyclopedia

Quellen, 2

- FAQ - UTF-8, UTF-16, UTF-32 & BOM
- UTF-8 - Wikipedia, the free encyclopedia
- UTF-16 - Wikipedia, the free encyclopedia
- ISO/IEC 8859-1 - Wikipedia, the free encyclopedia

Quellen, 3

- UAX #29: Unicode Text Segmentation
- Universal Character Set
- Unicode character property
- Comparison of Unicode encodings