Kódování znaků

Dříve existovaly teletypery/teleprintery - po síti byl poslán znak a tele-type ho natiskl na papír.

S příchodem počítačů vznikala spousta různých standardů. Například v Japonsku vznikly 4 různé standardy mezi sebou nekompatibilní. To při odeslání textu z jednoho počítače na druhý produkovalo jen rozházený text. V japonštině pro označení takové situace dokonce vzniklo slovo mojibake.

ASCII

American Standard Code for Information Interchange. Standard pro kódování znaků vytvořený primárně pro teleprintery. Znaky jsou kódovány v sedmi bitech. Počet možných znaků je tedy 128. Znaky mají přesně definované číslo, ke kterému patří.

Znak A je v ASCII definován jako 65 1000001 a znak a jako 97 1100001. Lze tak jednoduše určit zakódované znaky *okem* přímo z binární soustavy.

Dec	Hex	Char		Dec	Hex	Html	Char	r	Dec	Hex	Html	Char	Dec	Hex	Html	Char
Θ	θ	NUL	(NULL character)	32	20	6#32;	SP	(Space)	64	48	@		96	60	`	
1	1	SOH	(Start of Header)	33	21	6#33;	1.0		65	41	A	A	97	61	a	a
2	2	STX	(Start of Text)	34	22	6#34;			66	42	B	В	98	62	b	b
3	3	ETX	(End of Text)	35	23	#	#		67	43	C	C	99	63	c	c
4	4	E0T	(End of Transmission)	36	24	δ#36;	\$		68	44	D	D	100	64	d	d
5	5	ENQ	(Engu1ry)	37	25	6#37;	%		69	45	E	E	101	65	e	e
6	6	ACK	(Acknowledge)	38	26	6#38;	8		70	46	F	F	102	66	f	f
7	7	BEL	(Bell)	39	27	'			71	47	G	6	103		g	g
8	8	BS	(Backspace)	40	28	6#40;	(72	48	8#72;	H	104	68	h	h
9	9	HT	(Horizontal Tab)	41	29	6#41;)		73	49	I	I	105	69	i	i
10	Α		(Line feed)	42	2A	6#42;	*		74	4A	δ#74;		106	6A	j	j
11	В	VT	(Vertical Tab)	43	2B	6#43;	+		75	4B	δ#75;	K	107	6B	k	k
12	C	FF	(Form Feed)	44	2C	6#44;			76	4C	L	L L	108	6C	l	ι
13	D	CR	(Carriage return)	45	2D	8,445;	-		77	4D	δ#77;	M	109	6D	m	n
14	E	S0	(Shift Out)	46	2E	8#46;			78	4E	N	N	110	6E	n	n
15	F	SI	(Shift In)	47	2F	8#47;	1		79	4F	δ#79;	0	111	6F	o	0
16	10	DLE	(Data Link Escape)	48	30	6#48;	Θ		80	50	£#80;	P	112	70	p	P
17	11	DC1	(Device Control (XOn))	49	31	6#49;	1		81	51	Q	Q	113	71	q	q
18	12	DC2	(Device Control)	50	32	δ#50;	2		82	52	R	R	114	72	8#114;	r
19	13	DC3	(Device Control (XOff))	51	33	3	3		83	53	S	S	115	73	s	S
28	14	DC4	(Device Control)	52	34	6#52;	4		84	54	T	T	116	74	t	t
21	15		(Negative Acknowledge)	53	35	6#53;	5		85	55	U	U	117	75	u	u
22	16	SYN	(Synchronous Idle)	54	36	6#54;	6		86	56	V		118	76	v	V
23	17	ETB	(End of Transmission Block)	55	37	6#55;	7		87	57	δ#87;	W	119	77	w	W
24	18	CAN	(Cancel)	56	38	8#56;	8		88	58	X	X	120	78	x	×
25	19	EM	(End of Medium)	57	39	9	9		89	59	Y	Y	121	79	y	у
26	1A	SUB	(Substitute)	58	ЗА	:	1.0		90	5A	Z	Z	122	7A	z	Z
27	1B	ESC	(Escape)	59	3B	8#59;	- 1		91	5B	δ#91;	1	123	7B	6#123;	(
28	10	FS	(File Separator)	60	30	8#60;	<		92	5C	δ#92;	A. Comment	124	7C	8#124;	1
29	10	GS	(Group Separator)	61	3D	6#31;	-		93	5D	δ#93;	1	125	7D	}	}
38	1E	RS	(Record Separator)	62	3E	δ#62;	>		94	5E	δ#94;	^	126	7E	~	~
31	1F	US	(Unit Separator)	63	3F	?	7		95	5F	_		127	7F		DEL (Delete)

Obrázek 1: ASCII tabulka

Z ASCII později vycházely standardy pro specifické jazyky, které mezi sebou nejsou kompatibilní.

UTF-8

Standard prvotně načrtnutý na ubrousku v jídelně (Ken Thompson a Rob Pike). Unicode má seznam 277021 znaků ve verzi (11.0 červen 2018). UTF-8 může kódovat až 1112064 znaků.

Pokud v UTF-8 chceme zakódovat znak, který patří do ASCII tabulky (jde zakovat v sedmi bitech), zakódujeme ho stejně a na začátek přidáme nulu, tedy A 65 01000001.

Další znaky se kódují pomocí hlaviček v bytech.

volné bity	byte 0	byte 1	byte 2	byte 3
11	110xxxxx	10xxxxxx		
16	1110xxxx	10xxxxxx	10xxxxxx	
21	11110xxx	10xxxxxx	10xxxxxx	10xxxxxx

110 udává, kolik hlaviček v řetězci je. V dalším bajtu je hlavička 10, která označuje začátek bloku. Prázdná místa x jsou použita jako datová. Pomocí hlaviček získáme strukturu dat a jejich odstraněním z řetězce bitů získáme samotná data.

UTF-8	Obsah bloků bez hlaviček	Číslo v base10
110 10001 10 111000	10001111000	1114
1110 1001 10 1011110 10 011011	1001101110011011	39835

• Nenastane situace, kdy bude osm 0 za sebou

• Neplýtvá se místem