evropský sociální fond v ČR
EVROPSKÁ UNIE
MINISTERSTVO ŠKOLSTVÍ, MLÁDEŽE A TĚLOVÝCHOVY
OP Vzdělávání pro konkurenceschopnost
INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

PSK2-19

Název školy:	Vyšší odborná škola a Střední průmyslová škola, Božetěchova 3
Autor:	Ing. Marek Nožka
Anotace:	Kódování textu, znakové sady
Vzdělávací oblast:	Informační a komunikační technologie
Předmět:	Počítačové sítě a komunikační technika (PSK)
Tematická oblast:	Vrstvy protokolu TCP/IP
Výsledky vzdělávání:	Žák vysvětluje kódování textu a národních znaků
Klíčová slova:	ASCII, UTF-8, Unicode, národní znaková sada
Druh učebního materiálu:	Video-prezentace, Online vzdělávací materiál
Typ vzdělávání:	Střední vzdělávání, 3. ročník, technické lyceum
Ověřeno:	VOŠ a SPŠE Olomouc; Třída: 3L
Zdroj:	Vlastní poznámky, Wikipedia, Wikimedia Commons

Kódování textu

Text se v počítači kóduje pomocí tzv. znakové sady. Ta přiřazuje každému znaku určité číslo. Je to proto, že počítače dokáži pracovat pouze s čísly. Situace by se dala přirovnat k Morseově abecedě, jen s tím rozdílem, že zde nekódujeme pomocí teček a čárek, ale pomocí jedniček a nul.

ASCII

(ASCII je anglická zkratka pro American Standard Code for Information Interchange -- „americký standardní kód pro výměnu informací“).

Jedná se o znakovou znakovou zadu, která definuje znaky anglické abecedy, a jiné znaky používané v informatice. Jde o historicky nejúspěšnější znakovou sadu, z které vychází většina současných standardů pro kódování textu přinejmenším v euro-americké zóně.

V této znakové sadě odpovídá každý znak 7 respektive 8 bitům -- jednomu bajtu. To je velice výhodné pro programy, které s textem

pracují.

ASCII tabulka

Znaky a jejich číselné vyjádření je zapsáno v tzv. ASCII tabulce. Ta obsahuje

- tisknutelné znaky:
  - písmena, číslice,
  - jiné znaky (závorky, matematické znaky (+ - * / % atd.),
  - interpunkční znaménka (, . : ; atd.),
  - speciální znaky (@ $ ~atd.)),
- řídící (netisknutelné) kódy, které byly původně určeny pro řízení periferních zařízení (např. tiskárny nebo dálnopisu).

Dec	Hex		Dec	Hex		Dec	Hex		Dec	Hex		Dec	Hex		Dec	Hex		Dec	Hex	
0	00	NUL	16	10	DLE	32	20		48	30	0	64	40	@	80	50	P	96	60	`	112	70	p
1	01	SOH	17	11	DC1	33	21	!	49	31	1	65	41	A	81	51	Q	97	61	a	113	71	q
2	02	STX	18	12	DC2	34	22	"	50	32	2	66	42	B	82	52	R	98	62	b	114	72	r
3	03	ETX	19	13	DC3	35	23	#	51	33	3	67	43	C	83	53	S	99	63	c	115	73	s
4	04	EOT	20	14	DC4	36	24	$	52	34	4	68	44	D	84	54	T	100	64	d	116	74	t
5	05	ENQ	21	15	NAK	37	25	%	53	35	5	69	45	E	85	55	U	101	65	e	117	75	u
6	06	ACK	22	16	SYN	38	26	&	54	36	6	70	46	F	86	56	V	102	66	f	118	76	v
7	07	BEL	23	17	ETB	39	27	'	55	37	7	71	47	G	87	57	W	103	67	g	119	77	w
8	08	BS	24	18	CAN	40	28	(	56	38	8	72	48	H	88	58	X	104	68	h	120	78	x
9	09	HT	25	19	EM	41	29	)	57	39	9	73	49	I	89	59	Y	105	69	i	121	79	y
10	0A	LF	26	1A	SUB	42	2A	*	58	3A	:	74	4A	J	90	5A	Z	106	6A	j	122	7A	z
11	0B	VT	27	1B	ESC	43	2B	+	59	3B	;	75	4B	K	91	5B	[	107	6B	k	123	7B	{
12	0C	FF	28	1C	FS	44	2C	,	60	3C	<	76	4C	L	92	5C	\	108	6C	l	124	7C	|
13	0D	CR	29	1D	GS	45	2D	-	61	3D	=	77	4D	M	93	5D	]	109	6D	m	125	7D	}
14	0E	SO	30	1E	RS	46	2E	.	62	3E	>	78	4E	N	94	5E	^	110	6E	n	126	7E	~
15	0F	SI	31	1F	US	47	2F	/	63	3F	?	79	4F	O	95	5F	_	111	6F	o	127	7F	DEL

Konec řádku

Za zmínku stojí, že v různých operačních systémech se implementuje znak konce řádku různě:

- Nový řádek
- Carriage return
- Line feed

Národní znakové sady

Kód ASCII je podle původní definice sedmibitový (osmí bit byl původně paritní), obsahuje tedy **128 platných znaků**. Jsou to ale pouze znaky anglické abecedy. Pro potřeby dalších jazyků a pro rozšíření znakové sady se používají osmibitová rozšíření ASCII kódu, která obsahují dalších 128 znaků.

Takto rozšířený kód je přesto příliš malý na to, aby pojmul třeba jen evropské národní abecedy. Pro potřeby jednotlivých jazyků byly vytvořeny různé kódové tabulky, význam kódů nad 127 není tedy jednoznačný. Záleží na konkrétní národní znakové sadě. Takto vznikly rozdílné znakové sady například pro střední Evropu nebo pro Baltské jazyky.

V našem národním prostředí se nejvíce používá znaková sada ISO 8859-2 standardizovaná mezinárodní organizací pro normalizaci ISO nebo Windows-1250 (cp1250).

Unicode

Národní znakové sady ale nedostačují pokud chceme psát textový dokument a použít znaky z více jazyků. Pokud například chceme v českém textu použít znaky řecké abecedy je to ve znakové sadě ISO 8859-2 nebo cp1250 nemožné, protože tyto znakové sady obsahují znaky pro češtinu, slovenštinu, němčinu... ale neobsahují znaky pro řečtinu.

Unicode je tabulka znaků všech existujících abeced, která v současnosti obsahuje více než 110 000 znaků. Unicode umožňuje pracovat se znaky všech písem i různými jinými symboly stejným způsobem, takže mohou být využívány současně.

Každý znak má jednoznačný číselný kód a svůj název. Navíc Unicode definuje u každého znaku některé základní vlastnosti jako např. zda se jedná o písmeno, symbol atd., zda je písmeno velké či malé atp.

Původní návrh počítal s tím, že každý znak bude kódován 16-bitově, následně se ale ukázalo, že pro pokrytí všech používaných abeced to nestačí. V současné době, Tabulka Unicode poskytuje prostor pro 1.114.112 znaků s kódy 0_{HEX} až $10FFFF_{HEX}$. Tento prostor se dělí na 17 částí, každý o velikosti 2^{16}.

Existuje několik různých způsobů, jak znaky Unicode kódovat. Základní kódování, definovaná přímo ve standardu Unicode, jsou:

- UTF-32
- UTF-16
- UTF-8 -- UTF-8 je asi nejpoužívanější protože je zpětně kompatibilní s ASCII
- UCS-2

UTF-8

V UTF-8 se znaky kódují různě dlouhou (1–6 bajty) posloupností bajtů podle jejich pozice v Unicode. Znaky ASCII (`U+0000` —`U+007F`) jsou kódovány jedním bajtem, identicky jako v ASCII tím je dosaženo zpětné kompatibility s ASCII.

Znaky v rozsahu `U+0080` —`U+07FF`, kde jsou také všechny znaky s diakritikou používané v české abecedě jsou kódovány dvěma bajty

Znaky `U+0800` —`U+FFFF`, kam patří např znak Euro, €, `U+20AC` jsou kódovány třemi bajty, znaky mimo BMP jsou kódovány čtyřmi bajty.

Znaky, pro které by se použilo pětibajtové a šestibajtové kódování zatím nebyly definovány.

UTF-8 se často se používá pro přenos dat, neboť je prostorově úsporné (hlavně pro texty psané latinkou s nevelkým počtem znaků s diakritikou, které obsahují většinu jednobajtových a zbytek dvoubajtových kódů; v nelatinkových písmech je většina textu

tvořena dvoubajtovými kódy, písma Dálného východu používají tříbajtové kódy), je odolné proti chybám a zpětně kompatibilní s ASCII. Při jeho zpracování je však nepříjemná nestejná délka znaků.

Způsob kódování znaků

U+00000000 - U+0000007F	**0**xxxxxxx
U+00000080 - U+000007FF	**110**xxxxx 10xxxxxx
U+00000800 - U+0000FFFF	**1110**xxxx 10xxxxxx 10xxxxxx
U+00010000 - U+001FFFFF	**11110**xxx 10xxxxxx 10xxxxxx 10xxxxxx
U+00200000 - U+03FFFFFF	**111110**xx 10xxxxxx 10xxxxxx 10xxxxxx 10xxxxxx
U+04000000 - U+7FFFFFFF	**1111110**x 10xxxxxx 10xxxxxx 10xxxxxx 10xxxxxx 10xxxxxx