Представление текста

Кодировка текста

- Текст – последовательность букв.

  Если вам повезло родиться не в Китае.

- Память компьютера состоит из ячеек, в которых хранятся числа.

  Размер одной ячейки – 1 байт = 8 бит, может принимать значения от 0 до 255.

- Простейшим способом представления текста будет следующий:
  - для каждой буквы (и других символов) определим числовой код, так чтобы он помещался в 1 байт;
  - последовательность букв будем хранить как последовательность байт;
  - для того, чтобы понять где кончается текст можно использовать:
    - указание числа букв в тексте;
    - специальный код (например 0) в конце текста.

Кодировка ASCII

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

Кодовые страницы

- Для представления символов других алфавитов были придуманы кодовые страницы, использующие для обозначения дополнительных символов коды 128-255.
- Первые 127 символов совпадают с ASCII.
- Для русского языка использовались:
  - KOI-8 R (Unix)
  - CP866 (DOS)
  - CP1251 (Windows)
  - ISO 8859-5
- Для некоторых азиатских алфавитов использовалась кодировка DBCS (1 или 2 байта на символ)

Текст	H	e	l	l	o		П	р	и	в	е	т
Код ASCII	72	101	108	108	111	32						
CP866	72	101	108	108	111	32	143	224	168	162	165	226
CP1251	72	101	108	108	111	32	207	240	232	226	229	242

Путь символа на экран

Текст в исходной кодировке

- В тексте не написано, в какой именно

Текст в системной кодировке

- Символы, для которых нет аналога, заменяются на ?.

Символ на экране

- Символы, для которых в шрифте нет глифа, выводятся как □.