Кодировки символов

Кодировка символов – способ преобразования последовательности байтов в последовательность символов.

1. Однобайтовые кодировки

Один символ кодируется одним байтом

Максимально возможное количество символов в наборе символов – 256.

Во всех однобайтовых кодировках первая половина таблицы символов совпадает (00h-7Fh), но различается вторая половина таблицы символов (80h-FFh).

Примеры кодировок:

Windows-1251 (ANSI)

	00	01	02	03	04	05	06	07	08	09	0A	0B	0C	0D	0E	0F
00	NUL 0000	STX 0001	SOT 0002	ETX 0003	EOT 0004	ENQ 0005	ACK 0006	BEL 0007	BS 0008	HT 0009	LF 000A	VT 000B	FF 000C	CR 000D	SO 000E	SI 000F
10	DLE 0010	DC1 0011	DC2 0012	DC3 0013	DC4 0014	NAK 0015	SYN 0016	ETB 0017	CAN 0018	EM 0019	SUB 001A	ESC 001B	FS 001C	GS 001D	RS 001E	US 001F
20	SP 0020	! 0021	" 0022	# 0023	$ 0024	% 0025	& 0026	' 0027	(0028	) 0029	* 002A	+ 002B	, 002C	- 002D	. 002E	/ 002F
30	0 0030	1 0031	2 0032	3 0033	4 0034	5 0035	6 0036	7 0037	8 0038	9 0039	: 003A	; 003B	< 003C	= 003D	> 003E	? 003F
40	@ 0040	A 0041	B 0042	C 0043	D 0044	E 0045	F 0046	G 0047	H 0048	I 0049	J 004A	K 004B	L 004C	M 004D	N 004E	O 004F
50	P 0050	Q 0051	R 0052	S 0053	T 0054	U 0055	V 0056	W 0057	X 0058	Y 0059	Z 005A	[005B	\ 005C	] 005D	^ 005E	_ 005F
60	` 0060	a 0061	b 0062	c 0063	d 0064	e 0065	f 0066	g 0067	h 0068	i 0069	j 006A	k 006B	l 006C	m 006D	n 006E	o 006F
70	p 0070	q 0071	r 0072	s 0073	t 0074	u 0075	v 0076	w 0077	x 0078	y 0079	z 007A	{ 007B	| 007C	} 007D	~ 007E	DEL 007F
80	Ђ 0402	Ѓ 0403	‚ 201A	ѓ 0453	„ 201E	… 2026	† 2020	‡ 2021	€ 20AC	‰ 2030	Љ 0409	‹ 2039	Њ 040A	Ќ 040C	Ћ 040B	Џ 040F
90	ђ 0452	‘ 2018	’ 2019	“ 201C	” 201D	• 2022	– 2013	— 2014		™ 2122	љ 0459	› 203A	њ 045A	ќ 045C	ћ 045B	џ 045F
A0	NBSP 00A0	Ў 040E	ў 045E	Ј 0408	¤ 00A4	Ґ 0490	¦ 00A6	§ 00A7	Ё 0401	© 00A9	Є 0404	« 00AB	¬ 00AC	- 00AD	® 00AE	Ї 0407
B0	° 00B0	± 00B1	І 0406	і 0456	ґ 0491	µ 00B5	¶ 00B6	· 00B7	ё 0451	№ 2116	є 0454	» 00BB	ј 0458	Ѕ 0405	ѕ 0455	ї 0457
C0	А 0410	Б 0411	В 0412	Г 0413	Д 0414	Е 0415	Ж 0416	З 0417	И 0418	Й 0419	К 041A	Л 041B	М 041C	Н 041D	О 041E	П 041F
D0	Р 0420	С 0421	Т 0422	У 0423	Ф 0424	Х 0425	Ц 0426	Ч 0427	Ш 0428	Щ 0429	Ъ 042A	Ы 042B	Ь 042C	Э 042D	Ю 042E	Я 042F
E0	а 0430	б 0431	в 0432	г 0433	д 0434	е 0435	ж 0436	з 0437	и 0438	й 0439	к 043A	л 043B	м 043C	н 043D	о 043E	п 043F
F0	р 0440	с 0441	т 0442	у 0443	ф 0444	х 0445	ц 0446	ч 0447	ш 0448	щ 0449	ъ 044A	ы 044B	ь 044C	э 044D	ю 044E	я 044F

Windows-1251 используется в Windows по умолчанию. Кроме русского содержит символы большей части славянских языков.

CP866 (OEM)

	.0	.1	.2	.3	.4	.5	.6	.7	.8	.9	.A	.B	.C	.D	.E	.F
8.	А 410	Б 411	В 412	Г 413	Д 414	Е 415	Ж 416	З 417	И 418	Й 419	К 41A	Л 41B	М 41C	Н 41D	О 41E	П 41F
9.	Р 420	С 421	Т 422	У 423	Ф 424	Х 425	Ц 426	Ч 427	Ш 428	Щ 429	Ъ 42A	Ы 42B	Ь 42C	Э 42D	Ю 42E	Я 42F
A.	а 430	б 431	в 432	г 433	д 434	е 435	ж 436	з 437	и 438	й 439	к 43A	л 43B	м 43C	н 43D	о 43E	п 43F
B.	░ 2591	▒ 2592	▓ 2593	│ 2502	┤ 2524	╡ 2561	╢ 2562	╖ 2556	╕ 2555	╣ 2563	║ 2551	╗ 2557	╝ 255D	╜ 255C	╛ 255B	┐ 2510
C.	└ 2514	┴ 2534	┬ 252C	├ 251C	─ 2500	┼ 253C	╞ 255E	╟ 255F	╚ 255A	╔ 2554	╩ 2569	╦ 2566	╠ 2560	═ 2550	╬ 256C	╧ 2567
D.	╨ 2568	╤ 2564	╥ 2565	╙ 2559	╘ 2558	╒ 2552	╓ 2553	╫ 256B	╪ 256A	┘ 2518	┌ 250C	█ 2588	▄ 2584	▌ 258C	▐ 2590	▀ 2580
E.	р 440	с 441	т 442	у 443	ф 444	х 445	ц 446	ч 447	ш 448	щ 449	ъ 44A	ы 44B	ь 44C	э 44D	ю 44E	я 44F
F.	Ё 401	ё 451	Є 404	є 454	Ї 407	ї 457	Ў 40E	ў 45E	° B0	∙ 2219	· B7	√ 221A	№ 2116	¤ A4	■ 25A0	A0

Кодировка, которая используется в консольном режиме (эмуляция DOS). В Интернет используется редко. Кодировка создавалась таким образом, чтобы псевдографические символы, которые используются для рисования таблиц в DOS, были совместимы со стандартной западноевропейской кодировкой (чтобы псевдографика была одинаковой во всех кодировках).

Позиции русских символов не совпадают с Windows-1251, требуется конвертация текстов из одной кодировки в другую.

KOI8-R

	.0	.1	.2	.3	.4	.5	.6	.7	.8	.9	.A	.B	.C	.D	.E	.F
8.	─ 2500	│ 2502	┌ 250C	┐ 2510	└ 2514	┘ 2518	├ 251C	┤ 2524	┬ 252C	┴ 2534	┼ 253C	▀ 2580	▄ 2584	▉ 2588	▌ 258C	▐ 2590
9.	░ 2591	▒ 2592	▓ 2593	⌠ 2320	■ 25A0	· 2219	√ 221A	≈ 2248	≤ 2264	≥ 2265	A0	⌡ 2321	° B0	² B2	· B7	÷ F7
A.	═ 2550	║ 2551	╒ 2552	ё 451	╓ 2553	╔ 2554	╕ 2555	╖ 2556	╗ 2557	╘ 2558	╙ 2559	╚ 255A	╛ 255B	╜ 255C	╝ 255D	╞ 255E
B.	╟ 255F	╠ 2560	╡ 2561	Ё 401	╢ 2562	╣ 2563	╤ 2564	╥ 2565	╦ 2566	╧ 2567	╨ 2568	╩ 2569	╪ 256A	╫ 256B	╬ 256C	© A9
C.	ю 44E	а 430	б 431	ц 446	д 434	е 435	ф 444	г 433	х 445	и 438	й 439	к 43A	л 43B	м 43C	н 43D	о 43E
D.	п 43F	я 44F	р 440	с 441	т 442	у 443	ж 436	в 432	ь 44C	ы 44B	з 437	ш 448	э 44D	щ 449	ч 447	ъ 44A
E.	Ю 42E	А 410	Б 411	Ц 426	Д 414	Е 415	Ф 424	Г 413	Х 425	И 418	Й 419	К 41A	Л 41B	М 41C	Н 41D	О 41E
F.	П 41F	Я 42F	Р 420	С 421	Т 422	У 423	Ж 416	В 412	Ь 42C	Ы 42B	З 417	Ш 428	Э 42D	Щ 429	Ч 427	Ъ 42A

Кодировка создавалась для совместимости более старой 7-битовой кодировкой. Русские символы размещены таким образом, что если убрать старший бит байта, то русская буква превратится в латинскую, которая наиболее близка ей по звучанию (фонетически). Поэтому буквы следуют не в алфавитном порядке.

2. Кодировка Unicode (www.unicode.org)

Разрабатывается консорциумом «Юникод».
Для кодировки символа используется 4 байта (от 1 до 4 байтов).

Символы в кодировке Unicode соответствуют универсальному набору символов (Universal Character Set - UCS), он определен в ISO10646 (файл ISO_2003_charts_0000_33FF_BasicLatin_CJKCompat.pdf).

Латинские буквы и цифры – номера с 0000 по 007E
Русские буквы – номера с 0400 по 052F
(0400H = 0000 01XX XXXX XXXX)

В юникод существует возможность комбинации символов из нескольких, например, к символу какой-либо буквы можно добавить символ ударения.

В ссылках на символы (å) используются четырехбайтовые символы, определенные в ISO10646.

(Каждый символ занимает 4 байта и может не быть совместимости с однобайтовыми кодировками.)

Для решения проблемы в Unicode существует 3 формы кодирования: UTF-32, UTF-16, UTF-8

UTF-32
Для кодирования символа всегда используется 4 байта. Прямое соответствие номеру символа.
Наиболее простой вариант кодировки.

UTF-16
Если код символа содержит 16 бит (2 байта) то он кодируется напрямую, если 21 бит то с помощью четырех байтов. Коды символов в Unicode занимают не более 21 бита.
(файл ch03.pdf, стр.41)

Table 3-5. UTF-16 Bit Distribution

Scalar Value	UTF-16
xxxxxxxxxxxxxxxx	xxxxxxxxxxxxxxxx
000uuuuuxxxxxxxxxxxxxxxx	110110wwwwxxxxxx 110111xxxxxxxxxx

Note: wwww = uuuuu - 1

Поскольку большая часть символов занимает не более 16 бит, то по сравнению с UTF-32 получается выигрыш в размере данных (примерно в 2 раза).

UTF-8
Преимущество – кодировка первых 127 символов совпадает с набором символов ASCII (однобайтовой кодировкой). Символы латиницы не изменяются. Наиболее частый вариант кодировки.

Table 3-6. UTF-8 Bit Distribution

Scalar Value	First Byte	Second Byte	Third Byte	Fourth Byte
00000000 0xxxxxxx	0xxxxxxx			
00000yyy yyxxxxxx	110yyyyy	10xxxxxx		
zzzzyyyy yyxxxxxx	1110zzzz	10yyyyyy	10xxxxxx	
000uuuuu zzzzyyyy yyxxxxxx	11110uuu	10uuzzzz	10yyyyyy	10xxxxxx

Текстовые данные должны передаваться с использованием принятого в Интернет порядка байт ("big-endian", байт высшего порядка следует первым).
Также существует порядок "little-endian", байт низшего порядка следует первым.

В начале файла обычно располагаются несколько служебных байтов BOM – byte order mark, которые определяют big-endian или little-endian.

В UTF-8 BOM не имеет значения, так как все передается в виде байтов и их порядок передачи не важен. В 16 и 32 используются многобайтовые значения, поэтому важен порядок байтов и используется BOM.

Указание кодировки в HTML

Для задания кодировки в HTML используется элемент META (в заголовочной секции).

<META http-equiv="Content-Type" content="text/html; charset=КОДИРОВКА">

Этот элемент не производит перекодировку. Он только указывает браузеру в какой кодировке сохранен документ. Если кодировка документа не соответствует META (например документ в кодировке Windows-1251, а в META указана KOI8-R), то документ будет отображаться неправильно.