

Кодирование текстовой информации



Текстовая информация – информация, выраженная с помощью естественных и формальных языков в письменной форме.

Для кодирования букв русского алфавита
используется 8-разрядный код,
позволяющий закодировать – 256
различных символов.

Каждому символу ставится в соответствие –
уникальный десятичный код от 0 до 255
или

двоичный код от 00000000 до 11111111

Важно, что присвоение символу конкретного
кода – это вопрос соглашения, которое
фиксируется кодовой таблицей.

В качестве международного стандарта принята кодовая таблица **ASCII (American Standard Code for Information Interchange)**, кодирующая первую половину символов с числовыми кодами от 0 до 127 (коды от 0 до 32 отведены не символам, а функциональным клавишам).

Международная кодировка ASCII

sp 32	! 33	" 34	# 35	\$ 36	% 37	& 38	' 39	(40	) 41	* 42	+ 43	, 44	- 45	. 46	/ 47
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
@ 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
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
` 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
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	

Коды в кодовой таблице.

- ✓ **коды с 0 по 32** соответствуют не символам, а операциям (перевод строки, ввод пробела и т.д.).
- ✓ **коды с 33 по 127** являются интернациональными и соответствуют символам латинского алфавита, цифрам, знакам арифметических операций и знакам препинания.
- ✓ **коды с 128 по 255** являются национальными, т.е. в национальных кодировках одному и тому же коду соответствуют различные символы.

К сожалению, в настоящее время существует **5 различных кодовых страниц** для русских букв, поэтому тексты созданные в одной кодировке, не будут правильно отображаться в другой.

Стандарты кодировок:

1. КОИ-8 - UNIX
2. CP1251 («CP» означает «Code Page») - Microsoft Windows
3. CP 866 - MS-DOS
4. Mac - Macintosh
5. ISO 8859 – 5

Хронологически одним из первых стандартов кодирования русских букв на компьютерах был код **КОИ – 8** («**Код обмена информационный – 8 битный**»). Эта кодировка применяется в компьютерах с операционной системой UNIX.

Кодировка русских букв КОИ8-Р

—		Г	Г	└	└	┐	┐	┐	┐	┐	■	■	■	■	■
128	129	130	131	132	133	134	135	136	137	138	139	140	141	142	143
▤	▥	▦	┌	■	●	√	≈	≤	≥	nbsp	└	◦	²	•	÷
144	145	146	147	148	149	150	151	152	153	154	155	156	157	158	159
=		ƒ	ё	π	Г	Э	П	Г	Е	Ц	Ц	Г	Ц	Ц	Г
160	161	162	163	164	165	166	167	168	169	170	171	172	173	174	175
Г	Г	Г	Ё	Г	Г	Г	П	Г	Г	Ц	Ц	Г	Г	Г	©
176	177	178	179	180	181	182	183	184	185	186	187	188	189	190	191
Ю	А	Б	Ц	Д	Е	Ф	Г	Х	И	Й	К	Л	М	Н	О
192	193	194	195	196	197	198	199	200	201	202	203	204	205	206	207
П	Я	Р	С	Т	У	Ж	В	Ь	Ы	З	Ш	Э	Щ	Ч	Ъ
208	209	210	211	212	213	214	215	216	217	218	219	220	221	222	223
Ю	А	Б	Ц	Д	Е	Ф	Г	Х	И	Й	К	Л	М	Н	О
224	225	226	227	228	229	230	231	232	233	234	235	236	237	238	239
П	Я	Р	С	Т	У	Ж	В	Ь	Ы	З	Ш	Э	Щ	Ч	Ъ
240	241	242	243	244	245	246	247	248	249	250	251	252	253	254	255

Наиболее распространенная кодировка — это стандартная кириллическая кодировка Microsoft Windows, обозначаемая сокращением **CP1251** («CP» означает «Code Page»).

Все Windows – приложения, работающие с русским языком, поддерживают эту кодировку.

Кодировка русских букв CP1251

Á	à	,	è	„	...	†	‡	€	‰	É	<	й	Й	Ó	Ú
128	129	130	131	132	133	134	135	136	137	138	139	140	141	142	143
ă	‘	’	“	”	•	—	—	ě	™	é	>	ò	í	ó	ú
144	145	146	147	148	149	150	151	152	153	154	155	156	157	158	159
nbsp	ỳ	Ы	Э	Ѡ	Ы	!	§	Ё	©	Ю	«	¬	shy	®	Я
160	161	162	163	164	165	166	167	168	169	170	171	172	173	174	175
°	±	Ы	Э	’	μ	¶	•	ě	Nº	Ю	»	Э	Ю	Я	Я
176	177	178	179	180	181	182	183	184	185	186	187	188	189	190	191
А	Б	В	Г	Д	Е	Ж	З	И	Й	К	Л	М	Н	О	П
192	193	194	195	196	197	198	199	200	201	202	203	204	205	206	207
Р	С	Т	У	Ф	Х	Ц	Ч	Ш	Щ	Ъ	Ы	Ь	Э	Ю	Я
208	209	210	211	212	213	214	215	216	217	218	219	220	221	222	223
а	б	в	г	д	е	ж	з	и	й	к	л	м	н	о	п
224	225	226	227	228	229	230	231	232	233	234	235	236	237	238	239
р	с	т	у	ф	х	ц	ч	ш	щ	ъ	ы	ь	э	ю	я
240	241	242	243	244	245	246	247	248	249	250	251	252	253	254	255

Для работы в среде операционной системы MS-DOS используется «альтернативная» кодировка, в терминологии фирмы Microsoft — кодировка **CP 866**.

Кодировка русских букв CP866

А 128	Б 129	В 130	Г 131	Д 132	Е 133	Ж 134	З 135	И 136	Й 137	К 138	Л 139	М 140	Н 141	О 142	П 143
Р 144	С 145	Т 146	У 147	Ф 148	Х 149	Ц 150	Ч 151	Ш 152	Щ 153	Ъ 154	Ы 155	Ь 156	Э 157	Ю 158	Я 159
а 160	б 161	в 162	г 163	д 164	е 165	ж 166	з 167	и 168	й 169	к 170	л 171	м 172	н 173	о 174	п 175
 176	 177	 178	 179	├ 180	┤ 181	├┤ 182	П 183	└ 184	├┤ 185	 186	└┐ 187	┘┐ 188	┘ 189	└┐ 190	└ 191
└ 192	┐ 193	┐└ 194	└┐ 195	— 196	└┐ 197	└┐ 198	 199	└┐ 200	└┐ 201	└┐ 202	└┐ 203	└┐ 204	= 205	└┐ 206	└┐ 207
└┐ 208	└┐ 209	└┐ 210	└┐ 211	└┐ 212	└┐ 213	└┐ 214	└┐ 215	└┐ 216	└┐ 217	└┐ 218	■ 219	■ 220	■ 221	■ 222	■ 223
р 224	с 225	т 226	у 227	ф 228	х 229	ц 230	ч 231	ш 232	щ 233	ъ 234	ы 235	ь 236	э 237	ю 238	я 239
Ё 240	ё 241	Є 242	є 243	Ї 244	ї 245	Ў 246	ў 247	° 248	• 249	• 250	√ 251	Nº 252	¤ 253	■ 254	nbsp 255

Фирма Apple разработала для компьютеров Macintosh свою собственную кодировку русских букв (Mac)



Кодировка русских букв МАС

А	Б	В	Г	Д	Е	Ж	З	И	Й	К	Л	М	Н	О	П
128	129	130	131	132	133	134	135	136	137	138	139	140	141	142	143
Р	С	Т	У	Ф	Х	Ц	Ч	Ш	Щ	Ъ	Ы	Ь	Э	Ю	Я
144	145	146	147	148	149	150	151	152	153	154	155	156	157	158	159
†	°	Ы	£	§	•	¶	Ы	®	©	™	Á	á	ě	à	è
160	161	162	163	164	165	166	167	168	169	170	171	172	173	174	175
È	±	≤	≥	é	μ	г	Э	Ю	ю	Я	я	É	é	й	ò
176	177	178	179	180	181	182	183	184	185	186	187	188	189	190	191
Э	Ю	¬	√	f	≈	Δ	«	»	...	nbsp	Ó	ó	Й	й	я
192	193	194	195	196	197	198	199	200	201	202	203	204	205	206	207
–	—	“	”	‘	’	÷	„	ù	Ы	У	у	№	Ё	ё	я
208	209	210	211	212	213	214	215	216	217	218	219	220	221	222	223
а	б	в	г	д	е	ж	з	и	й	к	л	м	н	о	п
224	225	226	227	228	229	230	231	232	233	234	235	236	237	238	239
р	с	т	у	ф	х	ц	ч	ш	щ	ъ	ы	ь	э	ю	я
240	241	242	243	244	245	246	247	248	249	250	251	252	253	254	255



Международная организация по стандартизации (**International Standards Organization, ISO**) утвердила в качестве стандарта для русского языка еще одну кодировку под названием **ISO 8859 – 5**.

Кодировка русских букв ISO 8859-5

І	І	І	І	І	І	І	І	І	І	І	І	І	І	І	І
128	129	130	131	132	133	134	135	136	137	138	139	140	141	142	143
І	І	І	І	І	І	І	І	І	І	І	І	І	І	І	І
144	145	146	147	148	149	150	151	152	153	154	155	156	157	158	159
nbsp	Ё	Ъ	Ѓ	Є	Ѕ	І	Ї	Ј	Љ	Њ	Ћ	Ќ	shy	Ў	Џ
160	161	162	163	164	165	166	167	168	169	170	171	172	173	174	175
А	Б	В	Г	Д	Е	Ж	З	И	Й	К	Л	М	Н	О	П
176	177	178	179	180	181	182	183	184	185	186	187	188	189	190	191
Р	С	Т	У	Ф	Х	Ц	Ч	Ш	Щ	Ъ	Ы	Ь	Э	Ю	Я
192	193	194	195	196	197	198	199	200	201	202	203	204	205	206	207
а	б	в	г	д	е	ж	з	и	й	к	л	м	н	о	п
208	209	210	211	212	213	214	215	216	217	218	219	220	221	222	223
р	с	т	у	ф	х	ц	ч	ш	щ	ъ	ы	ь	э	ю	я
224	225	226	227	228	229	230	231	232	233	234	235	236	237	238	239
№	ё	ђ	ѓ	є	ѕ	і	ї	ј	љ	њ	ћ	ќ	ѕ	ў	џ
240	241	242	243	244	245	246	247	248	249	250	251	252	253	254	255

Таблица кодировки символов

Двоичный код	Десятичный код	КОИ8	CP1251	CP866	Mac	ISO
0000 0000	0					
.....						
0000 1000	8	Удаление последнего символа (клавиша Backspace)				
.....						
0000 1101	13	Перевод строки (клавиша Enter)				
.....						
0010 0000	32	Пробел				
0010 0001	33	!				
.....						
0101 1010	90	Z				
.....						
0111 1111	127					
.....	128	-	Ъ	А	А	К
.....						
1100 0010	194	Б	В	-	-	Т
.....						
1100 1100	204	Л	М	:	:	Ь
.....						19
1101 1101	221	Ш	Э	-	Ё	Н

В последнее время появился новый международный стандарт **Unicode**, который отводит на каждый символ не 8 символов, а 16, и поэтому с его помощью можно закодировать не 256 символов, $2^{16}=65\ 536$ различных символов.

Эту кодировку поддерживает платформа Microsoft Windows&Office97.

Последовательности десятичных кодов слова «ЭВМ»
в различных кодировках на основе кодировочных
таблиц:

КОИ8-Р	252	247	237
CP1251	221	194	204
CP866	157	130	140
Mac	157	130	140
ISO	205	178	188