Кодирование текстовой информации

Кодирование и декодирование

Хочется нам этого или нет, но современные компьютеры работают пока что только с числами. На экране могут быть и красивые картинки и веселое музыкальное сопровождение -- но для компьютера все это является лишь обработкой чисел. Любой объект, с которым работает компьютерная программа, должен быть так или иначе представлен в виде набора чисел, причем чисел двоичных.

Процесс представления любой информации в виде чисел называется *кодированием*.
Обратный процесс (из чисел -- в исходную информацию) -- *декодирование*.

Необходимость кодовых стандартов

Сосредоточимся на способах кодирования текста. Первый, напрашивающийся самим собой способ -- присвоить каждой букве из алфавита языка уникальное число.

Числа, поставленные в соответствие символам при таком способе кодирования, называются *кодами символов*.

Можем, например, договориться, что буква "А" у нас имеет код 1, "а" -- 2, "Б" -- 3, "б" -- 4, и т.д.

Таким образом, мы получаем таблицу соответствия символов и их кодов. Таблица эта называется *кодовой таблицей*.

Необходимость кодовых стандартов

Что же теперь делать с этой кодировкой? Обучать ей свои программы и периферийные устройства компьютера.

Можно сделать принтер так, чтобы, если на него поступает код 1, он изображал букву "А", а если 3 -- букву "Б". От клавиатуры также можно добиться того, что, если мы нажимаем на клавишу "А", в компьютер посылается код 1, а если "Б" -- 3.

Но это еще не все. Если Вася придумал себе одну кодировку, а Петя -- другую, пользы от этого будет немного. Ведь они не смогут свободно обмениваться текстами.

Кодировки у них разные.

Необходимость кодовых стандартов

Для того чтобы они смогли обмениваться текстами, нужно перед печатью преобразовать Васин текст в Петину кодировку. Для этого надо сообщить компьютеру, что число 1 из Васиного сочинения надо везде заменить на 45, число 3 -- на 46 и т.д. После этого уже можно будет смело отправлять сочинение на Петин принтер.

Этот процесс, процесс перевода закодированного текста из одной кодировки в другую, называется *перекодировкой текста*, а сама программа -- *перекодировщиком* или *конвертером*.

Необходимость кодовых стандартов

В случае с двумя пользователями дело обстоит просто. А если их сотня миллионов? Писать 2000000000000000 программ-конвертеров? Нет.

Куда проще придумать для всех пользователей одну-единственную, подходящую под нужды любого пользователя кодировку -- и заставить всех в мире пользоваться только ею. То есть, принять кодировочный стандарт.

8-разрядное кодирование. Система ASCII

(American Standard Code for Information Interchange)

8-разрядное кодирование заключается в том, что каждому символу ставится в соответствие уникальный двоичный код от 00000000 до 11111111 или соответствующий ему десятичный код от 0 до 255. То есть каждому символу отводится 1 байт памяти.

В системе ASCII закреплены две таблицы кодирования: базовая и расширенная.
Базовая закрепляет значения кодов от 0 до 127.
Расширенная таблица относится к символам с номерами от 128 до 255.

Базовая таблица ASCII

Десятичный код	Символ	Десятичный код	Символ	Десятичный код	Символ	Десятичный код	Символ
0		8	◘	16	►	24	↑
1	☺	9		17	◄	25	↓
2	☻	10		18	↕	26	›
3	♥	11	♂	19	‼	27	‹
4	♦	12	♀	20	¶	28	∟
5	♣	13		21	§	29	‹ ›
6	♠	14	♫	22	▬	30	▲
7	•	15	☼	23	↨	31	▼

Первые 32 кода базовой таблицы, начиная с нулевого, отданы производителям аппаратных средств (компьютеров, печатающих устройств). Это так называемые управляющие коды, которым не соответствуют никакие символы языков. Ими можно управлять выводом данных.

Базовая таблица ASCII

Десятичный код	Символ	Десятичный код	Символ	Десятичный код	Символ	Десятичный код	Символ
32		56	8	80	P	104	h
33	!	57	9	81	Q	105	i
34	"	58	:	82	R	106	j
35	#	59	;	83	S	107	k
36	$	60	<	84	T	108	l
37	%	61	=	85	U	109	m
38	&	62	>	86	V	110	n
39	‘	63	?	87	W	111	o
40	(	64	@	88	X	112	p
41	)	65	A	89	Y	113	q
42	*	66	B	90	Z	114	r
43	+	67	C	91	[	115	s
44	,	68	D	92	\	116	t
45	-	69	E	93	]	117	u
46	.	70	F	94	^	118	v
47	/	71	G	95	_	119	w
48	0	72	H	96	`	120	x
49	1	73	I	97	a	121	y
50	2	74	J	98	b	122	z
51	3	75	K	99	c	123	{
52	4	76	L	100	d	124	|
53	5	77	M	101	e	125	}
54	6	78	N	102	f	126	~
55	7	79	O	103	g	127	⌂

С 32 по 127 размещены коды символов латинского алфавита, знаков препинания, цифр, арифметических действий и специальных символов.

Расширенная таблица ASCII

Расширенная таблица относится к символам с номерами от 128 до 255. Здесь расположены национальные системы кодирования. Отсутствие единого стандарта в этой области привело к множественности одновременно действующих кодировок.

В настоящее время наиболее часто можно встретить следующие кодовые страницы для русских букв :

- Альтернативная кодировка, она же IBM CP866 — в системах DOS;
- Windows-1251, она же Microsoft code page 1251 (CP1251), в системах Windows;
- Семейство кодовых страниц KOI8 — в системах на основе UNIX (и Linux);

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

Альтернативная кодировка

Альтернативная кодировка – кодовая страница, где все специфические европейские символы во второй половине заменены на кириллицу, оставляя псевдографические символы нетронутыми. Следовательно, это не портит вид программ, использующих для работы текстовые окна, а также обеспечивает использование в них символов кириллицы. Альтернативная кодировка всё ещё жива и чрезвычайно популярна в среде DOS и OS/2. Кроме того, в этой кодировке записываются имена в файловой системе FAT (и короткие имена в VFAT). CP866 до сих пор используется в консоли русифицированных систем семейства Windows NT.

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

Кодировка Windows-1251

Windows-1251 —кодировка, являющаяся стандартной 8-битной кодировкой для всех русских версий Microsoft Windows. Пользуется довольно большой популярностью. Windows-1251 выгодно отличается от других 8-битных кириллических кодировок (таких как CP866, KOI8-R и ISO-8859-5) наличием практически всех символов, использующихся в русской типографике для обычного текста (отсутствует только значок ударения); она также содержит все символы для близких к русскому языку языков: украинского, белорусского, сербского и болгарского.

Имеет два недостатка:

• строчная буква я имеет код 255 в десятичной системе. Она является виновницей ряда неожиданных проблем в программах использующих этот код как служебный ;

• отсутствуют символы псевдографики

Десятичный код	Символ	Десятичный код	Символ	Десятичный код	Символ	Десятичный код	Символ
128	─	160	═	192	ю	224	Ю
129	│	161	║	193	а	225	А
130	┌	162	╒	194	б	226	Б
131	┐	163	ё	195	ц	227	Ц
132	└	164	╓	196	д	228	Д
133	┘	165	╔	197	е	229	Е
134	├	166	╕	198	ф	230	Ф
135	┤	167	╖	199	г	231	Г
136	┬	168	╗	200	х	232	Х
137	┴	169	╘	201	и	233	И
138	┼	170	╙	202	й	234	Й
139	▀	171	╚	203	к	235	К
140	▄	172	╛	204	л	236	Л
141	▌	173	╜	205	м	237	М
142	▐	174	╝	206	н	238	Н
143	▀	175	╞	207	о	239	О
144	▒	176	╟	208	п	240	П
145	▒	177	╠	209	я	241	Я
146	▓	178	╡	210	р	242	Р
147	⌠	179	Ё	211	с	243	С
148	■	180	╢	212	т	244	Т
149	·	181	╣	213	у	245	У
150	√	182	╤	214	ж	246	Ж
151	≈	183	╥	215	в	247	В
152	≤	184	╦	216	ь	248	Ь
153	≥	185	╧	217	ы	249	Ы
154		186	╨	218	з	250	З
155	⌡	187	╩	219	ш	251	Ш
156	°	188	╪	220	э	252	Э
157	²	189	┘	221	щ	253	Щ
158	·	190	╬	222	ч	254	Ч
159	÷	191	©	223	ъ	255	Ъ

Кодировка KOI8 (русская)

KOI8 —кодовая страница, разработанная для кодирования букв кириллических алфавитов. Разработчики КОИ-8 поместили символы русского алфавита в таблице таким образом, что позиции кириллических символов соответствуют их фонетическим аналогам в английском алфавите в базовой таблице. Это означает, что если в тексте, написанном в КОИ-8, убирать восьмой бит каждого символа (отнять 128), то получается читабельный текст, хотя он и написан латинскими символами. Например, слова “Русский Текст” превратились бы в “rUSSKIJ tEKST”. Как побочное следствие, символы кириллицы оказались расположены не в алфавитном порядке.

Существует несколько вариантов кодировки КОИ-8 для различных кириллических алфавитов. Русский алфавит описывается в кодировке KOI8-R, украинский — в KOI8-U.

KOI8-R стал фактически стандартом для русской кириллицы в юникс-подобных операционных системах и электронной почте.

Обратите внимание!

Цифры кодируются по стандарту ASCII в двух случаях – при вводе-выводе и когда они встречаются в тексте.

Если цифры участвуют в вычислениях, то осуществляется их преобразование в другой двоичных код (см. урок «представление чисел в компьютере»).

Возьмем число **57**.
При использовании в тексте каждая цифра будет представлена своим кодом в соответствии с таблицей ASCII. 5 имеет код 53, а 7 имеет код 55. В двоичной системе это 00110101 00110111.

При использовании в вычислениях, код этого числа будет получен по правилам перевода числа 57 в двоичную систему и это будет 00111001.

16-разрядное кодирование. Кодировка UNICODE

В связи с изобилием систем кодирования возникает задача перекодировки символов. Это неудобно. Но если увеличить число разрядов в два раза, то число кодируемых символов возрастет до $2^{16} = 65536$. Этого хватит на латинский алфавит, кириллицу, иврит, африканские и азиатские языки, различные специализированные символы: математические, экономические, технические и многое другое.

Такая система, основанная на 16-ти разрядном кодировании, получила название универсальной – UNICODE. Каждому символу в такой кодировке отводится 2 байта памяти.

Главный недостаток Unicode состоит в том, что все тексты в этой кодировке становятся в два раза длиннее.

В настоящее время стандарты ASCII и Unicode мирно сосуществуют.

Примеры

1. Запишите в двоичном и шестнадцатеричном коде английскую букву A.
2. Запишите в двоичном и шестнадцатеричном коде в альтернативной кодировке русскую букву а
3. Запишите в двоичном и шестнадцатеричном коде в кодировке Windows-1251 русскую букву Я
4. Запишите символ, который в базовой кодировке имеет двоичный код 00110000
5. Запишите символ, который в альтернативной кодировке имеет двоичный код 10011111
6. Запишите символ, который в кодировке Windows-1251 имеет двоичный код 11100000
7. Какое сообщение закодировано в кодировке Windows-1251: 00110101 00100000 11100001 11100000 11101011 11101011 11101110 11100010