

Представление символьной информации

Лекция №1

Кодируемые символы

1. Буквенно-цифровые знаки алфавитов.
2. Специальные знаки: пробел, скобки, знаки препинания, знаки операций и т.д.
3. Управляющие символы.

Наиболее распространенные способы кодирования символов

1. Использование кодировочной таблицы **ASCII**.
2. Использование стандарта кодирования символов **Unicode**.

ASCII

ASCII – American Standard Code for Information Interchange (американский стандартный код обмена информацией)

Введен в действие институтом стандартизации США (ANSI – American National Standard Institute) в 1963 году.

Первоначально предполагалось использование 7 бит кода.

Таблица ASCII

ASCII Code Chart

	0	1	2	3	4	5	6	7	8	9	A	B	C	D	E	F
0	NUL	SOH	STX	ETX	EOT	ENQ	ACK	BEL	BS	HT	LF	VT	FF	CR	SO	SI
1	DLE	DC1	DC2	DC3	DC4	NAK	SYN	ETB	CAN	EM	SUB	ESC	FS	GS	RS	US
2		!	"	#	\$	%	&	'	()	*	+	,	-	.	/
3	0	1	2	3	4	5	6	7	8	9	:	;	<	=	>	?
4	@	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
5	P	Q	R	S	T	U	V	W	X	Y	Z	[\]	^	_
6	`	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o
7	p	q	r	s	t	u	v	w	x	y	z	{		}	~	DEL

01000100 01001110 01010011

DNS

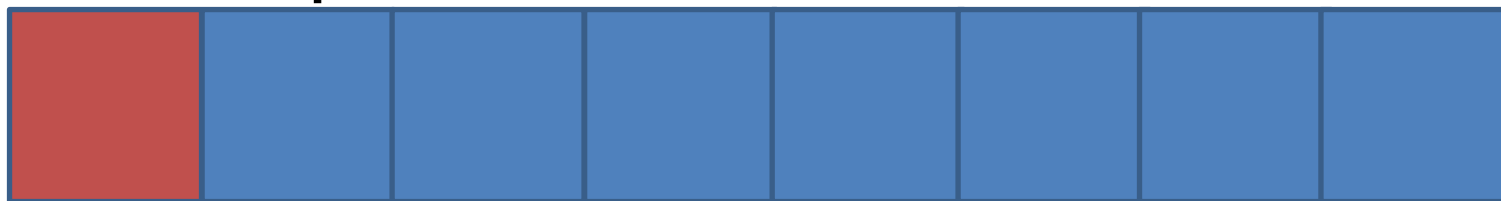
ЦИФРОВОЙ

01000100 01001110 01010011

Кодовая таблица ASCII

Кодовая таблица ASCII состоит из двух частей:

- Базовая таблица
- Расширенная таблица



Совокупность символов базовой и расширенной таблицы определяет *кодировку*.

Базовая таблица

ASCII Code Chart

	0	1	2	3	4	5	6	7	8	9	A	B	C	D	E	F
0	NUL	SOH	STX	ETX	EOT	ENQ	ACK	BEL	BS	HT	LF	VT	FF	CR	SO	SI
1	DLE	DC1	DC2	DC3	DC4	NAK	SYN	ETB	CAN	EM	SUB	ESC	FS	GS	RS	US
2		!	"	#	\$	%	&	'	()	*	+	,	-	.	/
3	0	1	2	3	4	5	6	7	8	9	:	;	<	=	>	?
4	@	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
5	P	Q	R	S	T	U	V	W	X	Y	Z	[\]	^	_
6	`	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o
7	p	q	r	s	t	u	v	w	x	y	z	{		}	~	DEL

Коды 00_h –
 $7F_h$

Расширенная таблица (ASCII)

128	Ç	144	É	160	á	176	░	192	Ł	208	⌌	224	α	240	≡
129	ü	145	æ	161	í	177	▒	193	⊥	209	⌍	225	β	241	±
130	é	146	Æ	162	ó	178	▓	194	⌒	210	⌎	226	Γ	242	≥
131	â	147	ô	163	ú	179		195	⌓	211	⌏	227	π	243	≤
132	ä	148	ö	164	ñ	180	┆	196	–	212	↳	228	Σ	244	∫
133	à	149	ò	165	Ñ	181	┆	197	+	213	ℱ	229	σ	245	∫
134	â	150	û	166	ª	182		198	⌔	214	ℙ	230	μ	246	÷
135	ç	151	ù	167	º	183	π	199	⌕	215	⌚	231	τ	247	≈
136	ê	152	ÿ	168	¸	184	ƒ	200	⌖	216	⌛	232	Φ	248	°
137	ë	153	Ö	169	¸	185		201	ℙ	217	⌜	233	⊙	249	.
138	è	154	Ü	170	¬	186		202	⌞	218	⌝	234	Ω	250	.
139	ì	155	◊	171	½	187	π	203	⌟	219	■	235	δ	251	√
140	î	156	£	172	¼	188	⌞	204	⌠	220	■	236	∞	252	π
141	ï	157	¥	173	¡	189	⌞	205	=	221	▬	237	φ	253	z
142	Ä	158	ℙ	174	«	190	⌞	206	⌡	222	▬	238	ε	254	■
143	Å	159	f	175	»	191	⌞	207	⌢	223	■	239	∩	255	

Source: www.LookupTables.com

Коды 80_h –
FF.

КОИ-8

	0	1	2	3	4	5	6R	7	8	9	A	B	C	D	E	F
80	2500 —	2502 	250C └	2510 └	2514 └	2518 └	251C └	2524 └	252C └	2534 └	253C └	2580 ■	2584 ■	2588 ■	258C ■	2590 ■
90	2591 ▒	2592 ▒	2593 ▒	2320 └	25A0 ■	2219 •	221A √	2248 ≈	2264 ≤	2265 ≥	A0 └	2321 └	B0 °	B2 2	B7 .	F7 ÷
A0	2550 =	2551 	2552 └	451 ë	2553 └	2554 └	2555 └	2556 └	2557 └	2558 └	2559 └	255A └	255B └	255C └	255D └	255E └
B0	255F └	2560 └	2561 └	401 Ë	2562 └	2563 └	2564 └	2565 └	2566 └	2567 └	2568 └	2569 └	256A └	256B └	256C └	A9 ©
C0	44E ю	430 а	431 б	446 ц	434 д	435 е	444 ф	433 г	445 х	438 и	439 й	43A к	43B л	43C м	43D н	43E о
D0	43F п	44F я	440 р	441 с	442 т	443 у	436 ж	432 в	44C ь	44B ы	437 з	448 ш	44D э	449 щ	447 ч	44A ъ
E0	42E Ю	410 А	411 Б	426 Ц	414 Д	415 Е	424 Ф	413 Г	425 Х	418 И	419 Й	41A К	41B Л	41C М	41D Н	41E О
F0	41F П	42F Я	420 Р	421 С	422 Т	423 У	416 Ж	412 В	42C Ь	42B Ы	417 З	428 Ш	42D Э	429 Щ	427 Ч	42A Ъ

Windows-1251 (CP 1251)

	0	1	2	3	4	5	6	7	8	9	A	B	C	D	E	F
80	402 Ъ	403 Ѓ	201A ,	453 ѓ	201E ,,	2026 ...	2020 †	2021 ‡	20AC €	2030 ‰	409 Љ	2039 ‹	40A Њ	40C Ќ	40B Ў	40F Џ
90	452 ђ	2018 ‘	2019 ’	201C “	201D ”	2022 •	2013 —	2014 —	□	2122 ™	459 љ	203A ›	45A њ	45C ќ	45B ћ	45F џ
A0	A0 	40E Ў	45E ў	408 Ј	A4 #	490 ѓ	A6 ;	A7 §	401 Ё	A9 ©	404 Є	AB «	AC ¬	AD -	AE ®	407 ї
B0	B0 °	B1 ±	406 І	456 і	491 ѓ	B5 µ	B6 ¶	B7 ·	451 ё	2116 №	454 є	BB »	458 ј	405 ѕ	455 ѕ	457 ї
C0	410 А	411 Б	412 В	413 Г	414 Д	415 Е	416 Ж	417 З	418 И	419 Й	41A К	41B Л	41C М	41D Н	41E О	41F П
D0	420 Р	421 С	422 Т	423 У	424 Ф	425 Х	426 Ц	427 Ч	428 Ш	429 Щ	42A Ъ	42B Ы	42C Ь	42D Э	42E Ю	42F Я
E0	430 а	431 б	432 в	433 г	434 д	435 е	436 ж	437 з	438 и	439 й	43A к	43B л	43C м	43D н	43E о	43F п
F0	440 р	441 с	442 т	443 у	444 ф	445 х	446 ц	447 ч	448 ш	449 щ	44A ъ	44B ы	44C ь	44D э	44E ю	44F я

Пример неправильно выбранной кодировки

Windows-1251 (CP 1251)

"Мой дядя самых честных правил,
Когда не в шутку занемог,
Он уважать себя заставил
И лучше выдумать не мог.
Его пример другим наука;
Но, боже мой, какая скука
С больным сидеть и день и ночь,
Не отходя ни шагу прочь!
Какое низкое коварство
Полуживого забавлять,
Ему подушки поправлять,
Печально подносить лекарство,
Вздыхать и думать про себя:
Когда же черт возьмет тебя!"

КОИ-8

Р
"лНИ дЪдЪ ЯЮЛШУ ВЕЯРМШУ ОПЮБХК
ЙНЦДЮ МЕ Б ЪСРЙС ГЮМЕЛНЦ,
НМ СВЮФЮРЭ ЯЕАЪ ГЮЯРЮБХК
Х КСВЪЕ ВШДСЛЮРЭ МЕ ЛНЦ.
еЦН ОПХЛЕП ДПСЦХЛ МЮСЙЮ;
мН, АНФЕ ЛНИ, ЙЮЙЮЪ ЯЙСЙЮ
я АНКЭМШЛ ЯХДЕРЭ Х ДЕМЭ Х МНВЭ
МЕ НРУНДЪ МХ БЮЦС ОПНВЭ!
ЙЮЙНЕ МХГЙНЕ ЙНБЮПЯРБН
ОНКСФХБНЦН ГЮАЮБКЪРЭ,
еЛС ОНДСЪЙХ ОНОПЮБКЪРЭ,
оЕВЮКЭМН ОНДМНЯХРЭ КЕЙЮПЯРБН,
БГДШУЮРЭ Х ДСЛЮРЭ ОПН ЯЕАЪ:
ЙНЦДЮ ФЕ ВЕПР БНГЭЛЕР РЕАЪ!"

Управляющие ASCII символы

Некоторые управляющие символы:

TAB, 09 - табуляция

LF, 0A - перевод строки

CR, 0D - возврат каретки

CR LF

Псевдографика

Для оформления программ и документов в текстовом режиме, используются *псевдографические символы*.

128	Ç	144	É	160	á	176	▤	192	⊥	208	⊥	224	α	240	≡
129	ù	145	æ	161	í	177	▥	193	⊥	209	⊥	225	β	241	±
130	é	146	Æ	162	ó	178	▧	194	⊥	210	⊥	226	Γ	242	≥
131	â	147	ô	163	ú	179		195	⊥	211	⊥	227	π	243	≤
132	ä	148	ö	164	ñ	180	⊥	196	—	212	⊥	228	Σ	244	∫
133	à	149	ò	165	Ñ	181	⊥	197	+	213	⊥	229	σ	245	∫
134	â	150	û	166	²	182	⊥	198	⊥	214	⊥	230	μ	246	+
135	ç	151	ù	167	°	183	⊥	199	⊥	215	⊥	231	τ	247	≈
136	ê	152	ÿ	168	¿	184	⊥	200	⊥	216	⊥	232	Φ	248	°
137	ë	153	Ö	169	Г	185	⊥	201	⊥	217	⊥	233	⊕	249	.
138	è	154	Û	170	Г	186	⊥	202	⊥	218	⊥	234	Ω	250	.
139	ï	155	◊	171	½	187	⊥	203	⊥	219	■	235	δ	251	√
140	î	156	£	172	¼	188	⊥	204	⊥	220	■	236	∞	252	∞
141	ï	157	¥	173	¡	189	⊥	205	=	221	■	237	φ	253	z
142	Ä	158	£	174	«	190	⊥	206	⊥	222	■	238	ε	254	■
143	Å	159	f	175	»	191	⊥	207	⊥	223	■	239	∩	255	

Unicode

Unicode – стандарт кодирования символов.

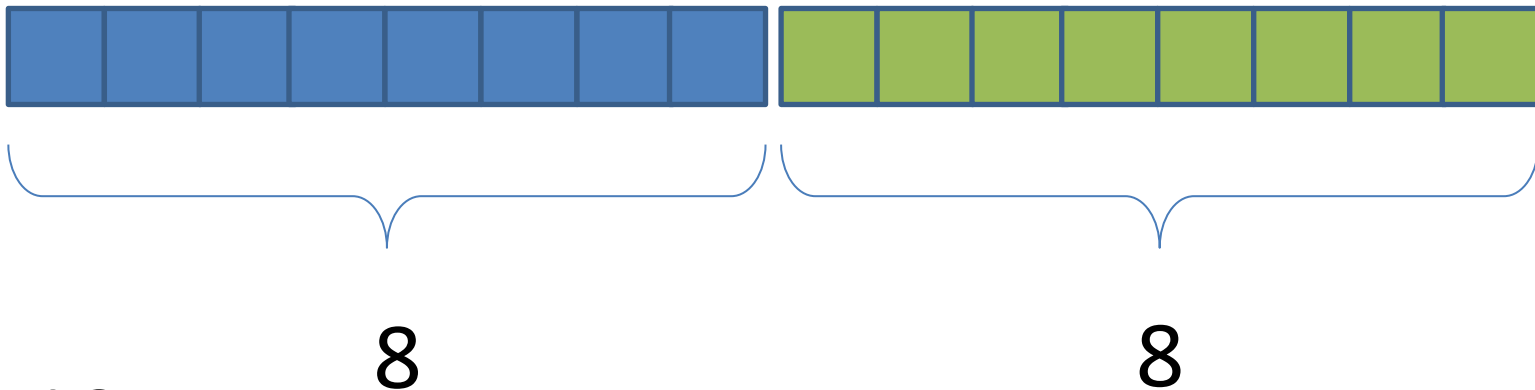
Коду символа сопоставляется некоторое положительное целое число.

Для представления кода в компьютере используются *форматы представления* (UTF – Unicode transformation format): UTF-8, UTF-16, UTF-32

По стандарту Unicode первые 128 символов соответствуют ASCII.

Unicode

В первых версиях стандарта код символа представлялся двухбайтовым словом



$$2^{16} = 65\,536$$

В настоящее время стандарт Unicode обеспечивает кодирование **1 112 064** символов.

UTF-8

8-битный формат преобразования Unicode

Обеспечивает совместимость с ASCII.

Длина кода нефиксированная – от 1 до 4 байт

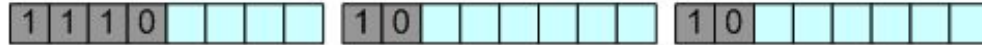
0x00000 – 0x0007F



0x00080 – 0x007FF



0x00800 – 0x0FFFF

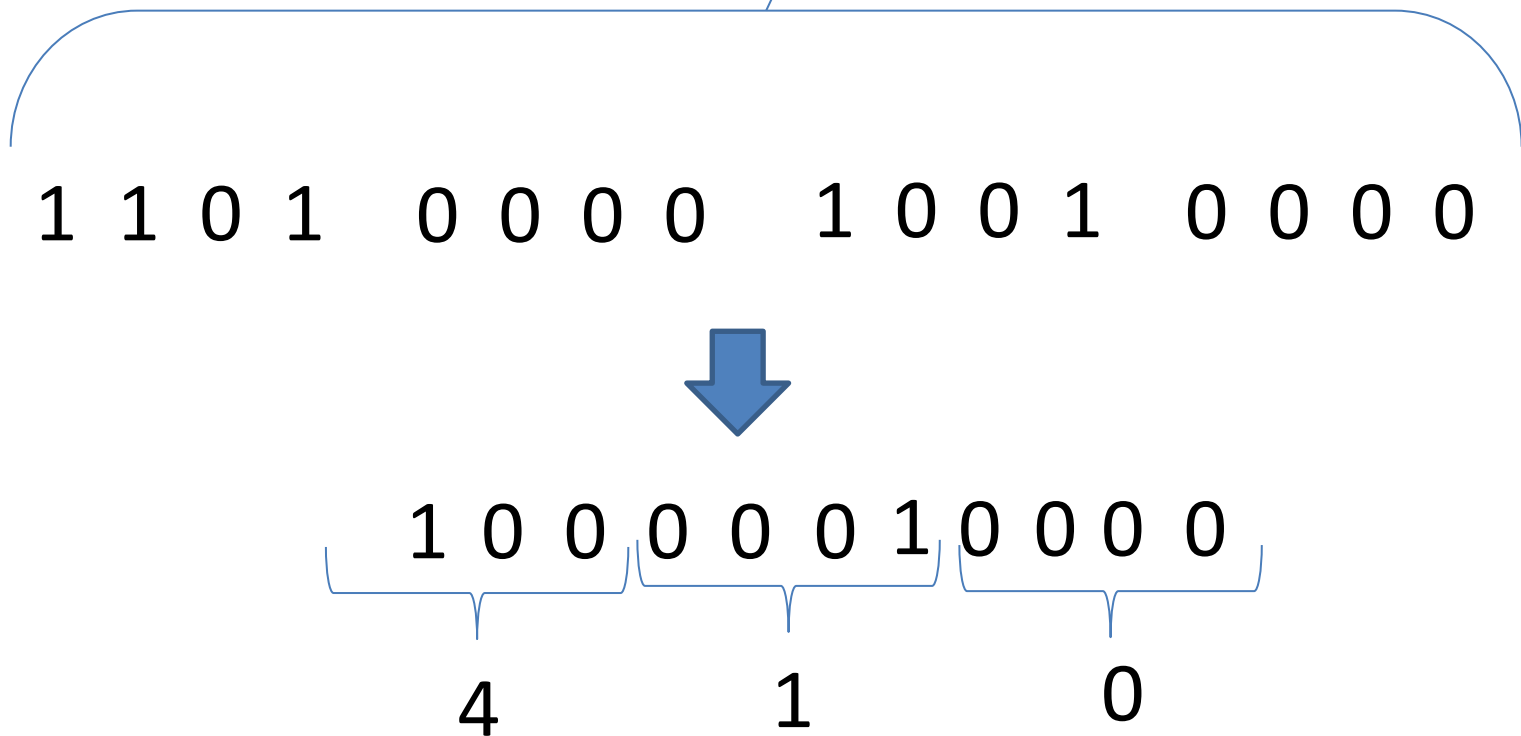


0x01000 – 0x1FFFF



Пример декодирования символа для UTF-8

Символ	UTF-8 (hex)	Unicode (hex)
А	D090	0410



Кириллическая таблица Unicode

	0	1	2	3	4	5	6	7	8	9	A	B	C	D	E	F
410	А	Б	В	Г	Д	Е	Ж	З	И	Й	К	Л	М	Н	О	П
420	Р	С	Т	У	Ф	Х	Ц	Ч	Ш	Щ	Ъ	Ы	Ь	Э	Ю	Я
430	а	б	в	г	д	е	ж	з	и	й	к	л	м	н	о	п
440	р	с	т	у	ф	х	ц	ч	ш	щ	ъ	ы	ь	э	ю	я

D0 BB D0 B5 D1 81

BOM

Byte Order Mark – сигнатура, определяющая UTF.

UTF-8

EF BB BF

UTF-16BE

FE FF

UTF-16LE

FF FE

UTF-32BE

00 00 FE FF

UTF-32LE

FF FE 00 00