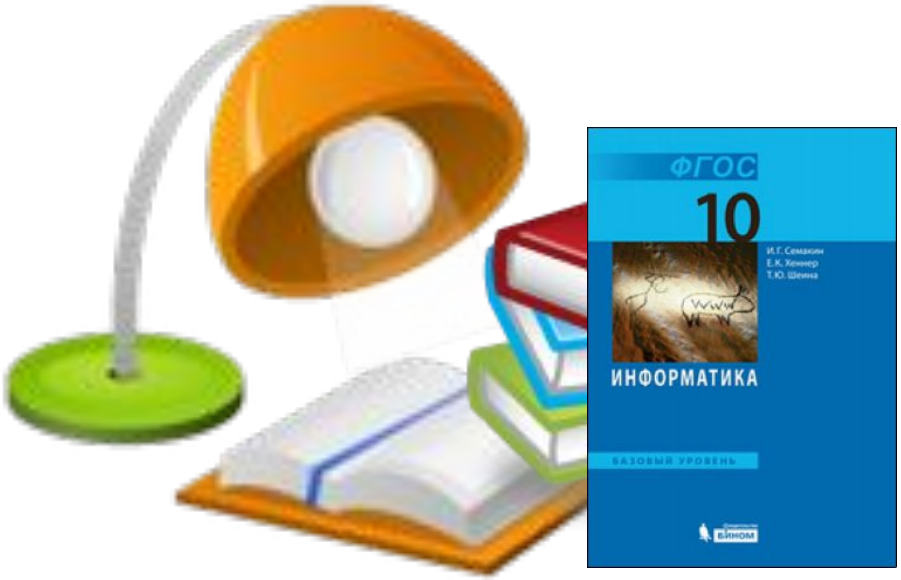
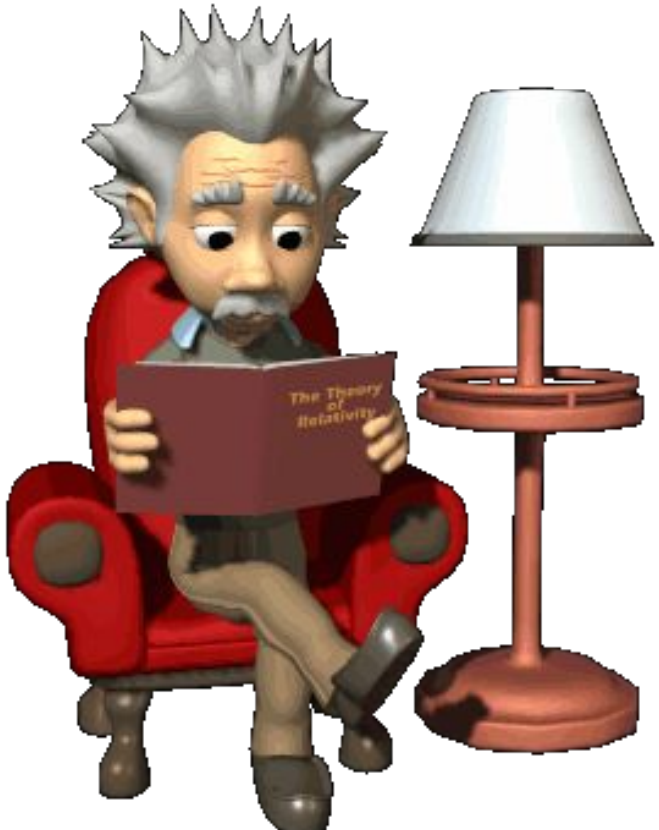
Кодирование текста

Урок 13

Домашнее задание

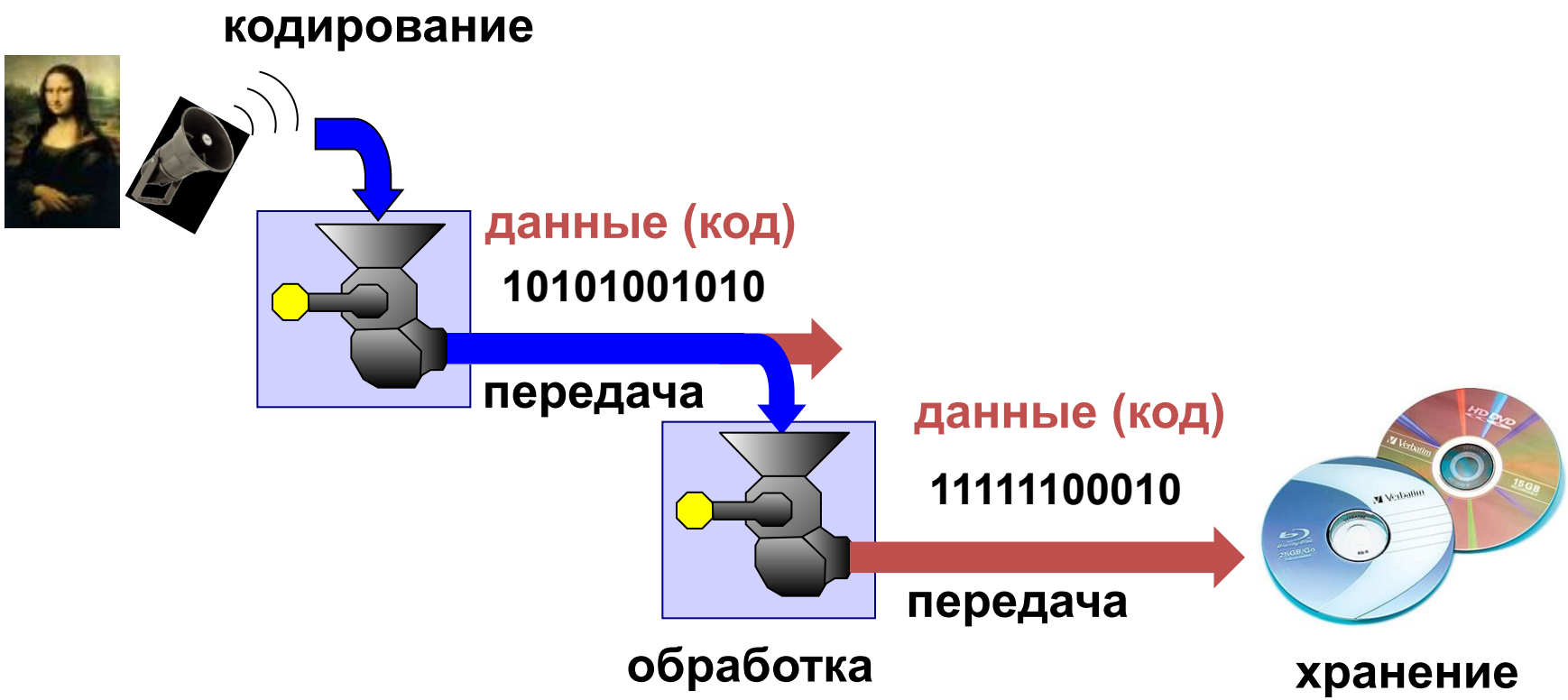
§6 (**начало** и **п1**) (стр.43–45) – выучить.
Вопрос **1** (стр. 51) – **устно**.





Представление данных и программ в компьютере

Итак, чтобы компьютер мог воспринять и обработать числовые значения, текст, изображение, звук или видео, их нужно представить в виде последовательностей 0 и 1



Кодирование текста

- на экране – **символы**
- в памяти – **двоичные коды**

65	66	67	68
1000001_2	1000010_2	1000011_2	1000100_2

! **В файле хранятся не изображения символов, а коды их порядковых номеров в двоичной системе!**

Вспомним

Если с помощью **n-разрядного** двоичного кода закодировать алфавит, то количество символов этого алфавита составит

$$N=2^n$$

n – информационный вес символа – количество бит в двоичном коде.
N – мощность алфавита – количество всех символов алфавита (кодовых комбинаций).

Кодовые таблицы

Для представления текстовых данных в компьютерах используют так называемые **кодовые таблицы** – наборы кодов для кодирования определенного количества символов, где каждому из символов соответствует двоичный код определенной длины.

A	01000001
B	01000010
C	01000011
D	01000100
E	01000101
F	01000110
G	01000111
H	01001000
I	01001001
J	01001010

Кодовая таблица ASCII

ASCII (англ. **American standard code for information interchange**, **[’æs.ki]**) — самая популярная кодовая таблица, была разработана и стандартизована в США в 1963 году. Название «ASCII» по-русски часто произносится как **[аски]**. Информационный вес символа в коде ASCII – **8** бит. Мощность алфавита при этом составляет **256** символов (**2^8**).

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

Первая половина таблицы ASCII

символ	10-й код	2-й код	символ	10-й код	2-й код	символ	10-й код	2-й код	символ	10-й код	2-й код
	32	00100000	8	56	00111000	P	80	01010000	h	104	01101000
!	33	00100001	9	57	00111001	Q	81	01010001	i	105	01101001
"	34	00100010	:	58	00111010	R	82	01010010	j	106	01101010
#	35	00100011	;	59	00111011	S	83	01010011	k	107	01101011
$	36	00100100	<	60	00111100	T	84	01010100	l	108	01101100
%	37	00100101	=	61	00111101	U	85	01010101	m	109	01101101
&	38	00100110	>	62	00111110	V	86	01010110	n	110	01101110
'	39	00100111	?	63	00111111	W	87	01010111	o	111	01101111
(	40	00101000	@	64	01000000	X	88	01011000	p	112	01110000
)	41	00101001	A	65	01000001	Y	89	01011001	q	113	01110001
*	42	00101010	B	66	01000010	Z	90	01011010	r	114	01110010
+	43	00101011	C	67	01000011	[	91	01011011	s	115	01110011
,	44	00101100	D	68	01000100	\	92	01011100	t	116	01110100
-	45	00101101	E	69	01000101	]	93	01011101	u	117	01110101
.	46	00101110	F	70	01000110	^	94	01011110	v	118	01110110
/	47	00101111	G	71	01000111	_	95	01011111	w	119	01110111
0	48	00110000	H	72	01001000	`	96	01100000	x	120	01111000
1	49	00110001	I	73	01001001	a	97	01100001	y	121	01111001
2	50	00110010	J	74	01001010	b	98	01100010	z	122	01111010
3	51	00110011	K	75	01001011	c	99	01100011	{	123	01111011
4	52	00110100	L	76	01001100	d	100	01100100	|	124	01111100
5	53	00110101	M	77	01001101	e	101	01100101	}	125	01111101
6	54	00110110	N	78	01001110	f	102	01100110	~	126	01111110
7	55	00110111	O	79	01001111	g	103	01100111	□	127	01111111

Вторая половина таблицы ASCII

символ	10-й код	2-й код	символ	10-й код	2-й код	символ	10-й код	2-й код	символ	10-й код	2-й код
Ђ	128	10000000		160	10100000	А	192	11000000	а	224	11100000
Ѓ	129	10000001	Ў	161	10100001	Б	193	11000001	б	225	11100001
‚	130	10000010	ў	162	10100010	В	194	11000010	в	226	11100010
ѓ	131	10000011	Ј	163	10100011	Г	195	11000011	г	227	11100011
„	132	10000100	¤	164	10100100	Д	196	11000100	д	228	11100100
…	133	10000101	Ґ	165	10100101	Е	197	11000101	е	229	11100101
†	134	10000110	¦	166	10100110	Ж	198	11000110	ж	230	11100110
‡	135	10000111	§	167	10100111	З	199	11000111	з	231	11100111
€	136	10001000	Ё	168	10101000	И	200	11001000	и	232	11101000
‰	137	10001001	©	169	10101001	Й	201	11001001	й	233	11101001
Љ	138	10001010	Є	170	10101010	К	202	11001010	к	234	11101010
‹	139	10001011	«	171	10101011	Л	203	11001011	л	235	11101011
Њ	140	10001100	¬	172	10101100	М	204	11001100	м	236	11101100
Ќ	141	10001101	-	173	10101101	Н	205	11001101	н	237	11101101
Ћ	142	10001110	®	174	10101110	О	206	11001110	о	238	11101110
Џ	143	10001111	Ї	175	10101111	П	207	11001111	п	239	11101111
ђ	144	10010000	°	176	10110000	Р	208	11010000	р	240	11110000
‘	145	10010001	±	177	10110001	С	209	11010001	с	241	11110001
’	146	10010010	І	178	10110010	Т	210	11010010	т	242	11110010
“	147	10010011	і	179	10110011	У	211	11010011	у	243	11110011
”	148	10010100	ґ	180	10110100	Ф	212	11010100	ф	244	11110100
•	149	10010101	µ	181	10110101	Х	213	11010101	х	245	11110101
–	150	10010110	¶	182	10110110	Ц	214	11010110	ц	246	11110110
—	151	10010111	·	183	10110111	Ч	215	11010111	ч	247	11110111
	152	10011000	ё	184	10111000	Ш	216	11011000	ш	248	11111000
™	153	10011001	№	185	10111001	Щ	217	11011001	щ	249	11111001
љ	154	10011010	є	186	10111010	Ъ	218	11011010	ъ	250	11111010
›	155	10011011	»	187	10111011	Ы	219	11011011	ы	251	11111011
њ	156	10011100	ј	188	10111100	Ь	220	11011100	ь	252	11111100
ќ	157	10011101	Ѕ	189	10111101	Э	221	11011101	э	253	11111101
ћ	158	10011110	ѕ	190	10111110	Ю	222	11011110	ю	254	11111110
џ	159	10011111	ї	191	10111111	Я	223	11011111	я	255	11111111

Проблема ASCII

Исторически сложилось, что в 8-битовых кодировках ASCII **первую половину** кодовой таблицы (0—127) занимают всегда **«американские» символы**,
а **вторую** (128—255) — дополнительные символы, включая набор букв национальных языков и местных символов.
Отсутствие единого стандарта размещения кириллических символов в таблице ASCII доставляло (и доставляет) множество проблем с кодировками (КОИ-8, Windows-1251 и др.).
Позже кодовые таблицы стандартизировали. Просто стандартизировали их названия и набор символов. Но проблема осталась!

Кодировка	Другие названия	Описание
ISO-8859-1		Западно-европейская Latin-1
CP1252	Windows-1252, 1252	Западно-европейская кодировка, применяемая в Windows.
CP866	DOS, 866	Кириллическая кодировка, применяемая в командном языке Windows.
CP1251	Windows-1251, win-1251, 1251, Windows	Кириллическая кодировка, применяемая в основном интерфейсе Windows.
KOI8-R	koi8r, koi8-ru	Русская кодировка. Поддерживается в ОС Unix.
BIG5	CP950, 950	Традиционная китайская, применяется в основном на Тайване.
GB2312	CP936, 936	Упрощенный китайская, стандартная национальная кодировка.
BIG5-HKSCS		Расширенная Big5, применяемая в Гонг-Конге.
Shift_JIS	CP932, SJIS, 932	Японская кодировка.
EUC-JP	EUCJP	Японская кодировка.

Кириллица в ASCII

К сожалению, в настоящее время существуют **много** различных кодовых таблиц для кириллицы в **ASCII**. Наиболее распространены **КОИ8-R**, **CP1251**, **CP866**, **Mac** и **ISO**. Из-за этого часто возникают проблемы с переносом русского текста с одного компьютера на другой, из одной программной системы в другую.

Разные кодировки кириллицы

Одним из первых стандартов кодирования русских букв был **КОИ8** ("Код обмена информацией, 8-битный"). Кодировка применялась ещё в 70-ые годы на компьютерах серии ЕС ЭВМ, а с середины 80-х годов стала использоваться в первых русифицированных версиях ОС **UNIX**. В дальнейшем используется «потомками» ОС Unix: **Linux**, **Android**.

От начала 90-х годов, времени господства операционной системы MS DOS, **остается** кодировка **CP866**. Используется в командном языке и в консольном режиме ОС **Windows**.

Наиболее распространенной в настоящее время является кодировка Microsoft, обозначаемая сокращением **CP1251**. Является стандартной 8-битной кодировкой для русских версий ОС **Windows**.

Компьютеры фирмы Apple, работающие под управлением операционной системы **Mac OS**, используют свою собственную кодировку **Mac**.

Кроме того, Международная организация по стандартизации (International Standards Organization, ISO) утвердила в качестве стандарта для русского языка еще одну кодировку под названием **ISO 8859-5**. Широко применяется в Сербии, Болгарии на юниксоподобных системах. У нас не популярна!

Unicode

С конца 90-х годов проблема стандартизации символьного кодирования решается введением нового международного стандарта, который называется **Unicode**. Это **16-разрядная** кодировка, т.е. в ней на каждый символ отводится **2 байта** памяти. Конечно, при этом объем занимаемой памяти увеличивается в 2 раза. Но зато такая кодовая таблица допускает включение **до 65536 символов**.

Полная спецификация стандарта **Unicode** включает в себя все существующие, вымершие и искусственно созданные алфавиты мира, а также множество математических, музыкальных, химических и прочих символов.

UTF-8

UTF-8 (от англ. **Unicode Transformation Format** — «формат преобразования Юникода, 8-битный») — одна из общепринятых и стандартизированных кодировок текста, которая позволяет хранить **символы Юникода**, используя **переменное количество байт** (от 1 до 6).
Коды символов первой половины кода **ASCII** совпадают с кодами **UTF-8**. Коды остальных символов содержат от 2 до 6 байт. Русские буквы – по 2 байта.

Разнообразие кодовых таблиц

В настоящее время наиболее распространенными кодами символов являются
ASCII – **8**-битный код,
Unicode – **16**-битный код,
UTF-8 – код с переменной длинной
и др.