

Кодирование текстовой информации



Кодирование текстов:

- Текст в памяти компьютера хранится как последовательность двоичных кодов символов. Изображения символов хранятся в отдельных шрифтовых файлах (расширения: **fon**, **ttf**, **otf**).
- Для предоставления текста в памяти компьютера необходимо представить его в виде двоичного код, т.е. в виде последовательности из **0** и **1**



Кодирование

- это процесс преобразования информации в форму, воспринимаемую компьютером, т.е. двоичный код.

Декодирование

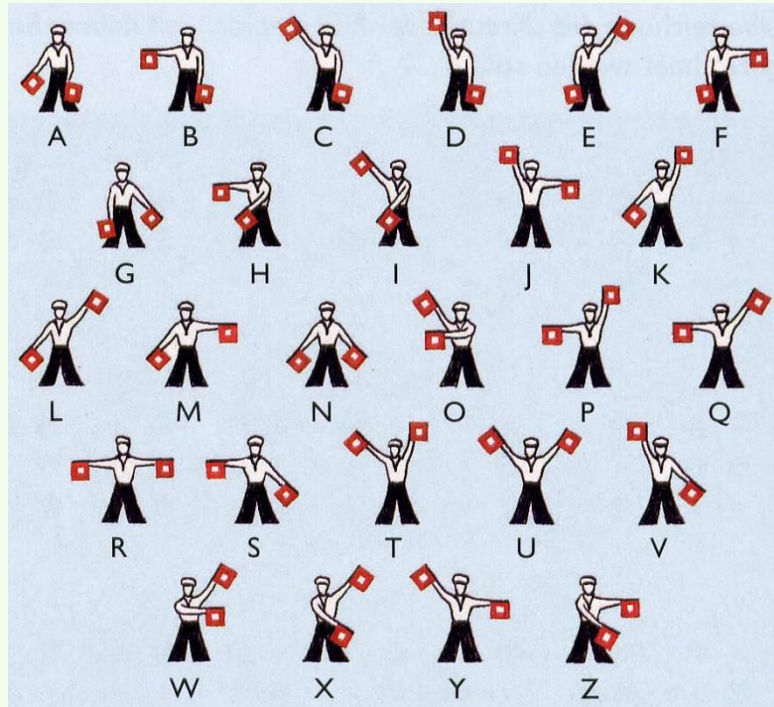
- это процесс обратный кодированию, который заключается в том, что происходит преобразование данных из двоичного кода в форму, понятную человеку.



Сигнальные коды

A	. -	S	. . .
B	- . . .	T	-
C	- . - .	U	. . -
D	- . .	V	. . . -
E	.	W	. - -
F	. . - .	X	- . . -
G	- - .	Y	- . - -
H		Z	- - . .
I	. .	1	. - - - -
J	. - - -	2	. . - - -
K	- . -	3	. . . - -
L	. - . .	4	 -
M	- -	5	
N	- .	6	-
O	- - -	7	- - . . .
P	. - - .	8	- - - . .
Q	- - - -	9	- - - - .
R	. - .	0	- - - - -

Азбука Морзе (1837 г)



Флажковая азбука

Способы кодирования

1) Замена буквы её порядковым номером в алфавите:

- $A \rightarrow 1, B \rightarrow 2, V \rightarrow 3, \dots, Я \rightarrow 33$

Замена буквы следующей за ней по алфавиту:

- $A \rightarrow B, B \rightarrow B, B \rightarrow Г, \dots, Я \rightarrow A$

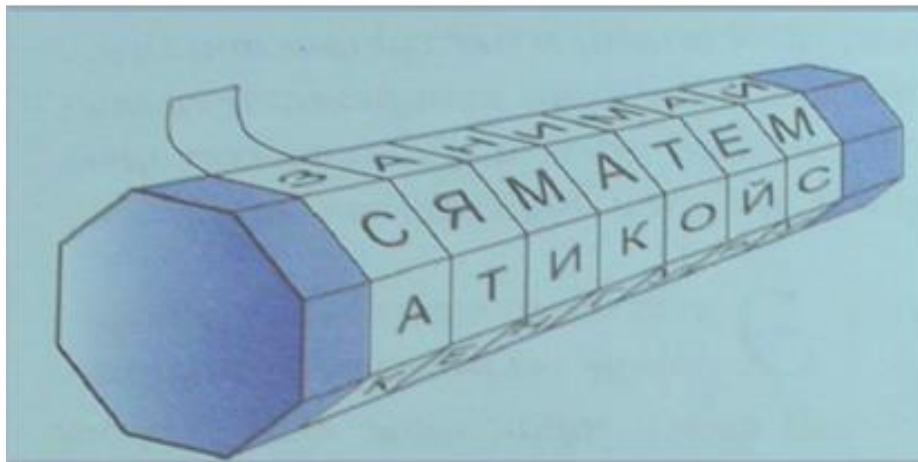
Замена буквы на букву, номер которой на 2 больше исходного:

- $A \rightarrow B, B \rightarrow Г, B \rightarrow Д, \dots, Я \rightarrow Б$



Шифр Сцитала

- Для прочтения шифровки нужно не только знать систему засекречивания, но и обладать ключом в виде палочки, принятого диаметра. Зная тип шифра, но, не имея ключа, расшифровать сообщение было сложно. Этот шифр именовался сцитала по названию стержня, на который наматывались свитки папируса, что указывает на его происхождение. Он был весьма популярен в Спарте и много раз совершенствовался в позднейшие времена



Шифр Энея

- **«диск Энея»** - небольшой диск, с просверленными вдоль его края отверстиями соответствующими буквам алфавита. Для того, чтобы зашифровать текст «шишка», нить вдевали в отверстия так, чтобы по ходу нити последовательность отверстий была такой же, как у соответствующих букв в тексте.
- Получатель сообщения, вытягивая эту нить, мог прочитать все буквы, правда, в обратном порядке.



Шифр Цезаря

- Шифр Цезаря предельно прост: в нем каждая буква сообщения заменялась на следующую за ней по алфавиту с указанием сдвига

А	Б	В	Г	Д	Е	Ж	З	И	Й	К	Л	М	Н	О	П	Р	С	Т	...	Ь	Ъ	Э	Ю	Я
Г	Д	Е	Ж	З	И	Й	К	Л	М	Н	О	П	Р	С	Т	У	Ф	Х		Ю	Я	А	Б	В

3



Шифр Вижинера

- При шифровании исходного сообщения его выписывают в строку, а под ним записывают ключевое слово. Если ключ оказался короче сообщения, то его циклически повторяют. В процессе шифрования очередная буква шифр-текста находится на пересечении столбца, определяемого шифруемой буквой, и строки, определяемой значением ключа.
- Сообщение П Р И Л Е Т А Ю С Е Д Ь М О Г О
- Ключ А М Б Р О З И Я А М Б Р О З И Я
- Шифр-текст П Ъ Й Ы У Щ И Э С С Е К Ъ Х Л Н



Таблица кодировки

Таблица, в которой всем символам компьютерного алфавита поставлены в соответствие порядковые номера (коды)

Международным стандартом для ПК стала таблица ASCII (Американский стандартный код для информационного обмена).

32 пробел	48 0	64 @	80 P	96 `	112 p
33 !	49 1	65 A	81 Q	97 a	113 q
34 "	50 2	66 B	82 R	98 b	114 r
35 #	51 3	67 C	83 S	99 c	115 s
36 \$	52 4	68 D	84 T	100 d	116 t
37 %	53 5	69 E	85 U	101 e	117 u
38 &	54 6	70 F	86 V	102 f	118 v
39 '	55 7	71 G	87 W	103 g	119 w
40 (	56 8	72 H	88 X	104 h	120 x
41)	57 9	73 I	89 Y	105 i	121 y
42 *	58 :	74 J	90 Z	106 j	122 z
43 +	59 ;	75 K	91 [	107 k	123 {
44 ,	60 <	76 L	92 \	108 l	124
45 -	61 =	77 M	93]	109 m	125 }
46 .	62 >	78 N	94 ^	110 n	126 ~
47 /	63 ?	79 O	95 _	111 o	127

Первую половину кодовой страницы (коды от **0** до **127**) занимает **стандартная** таблица ASCII, а **вторую** (коды от **128** до **256**) — **символы национальных алфавитов** (например, русские буквы).

Для русского языка существуют несколько кодовых страниц, которые были разработаны для разных операционных систем

в системе Windows

[illegible]

Windows-1251 (CP-1251)

в Unix-совместимых
операционных
системах и
электронной почте

128	144	160	176	192	208	224	240
129	145	161	177	193	209	225	241
130	146	162	178	194	210	226	242
131	147	163	179	195	211	227	243
132	148	164	180	196	212	228	244
133	149	165	181	197	213	229	245
134	150	166	182	198	214	230	246
135	151	167	183	199	215	231	247
136	152	168	184	200	216	232	248
137	153	169	185	201	217	233	249
138	154	170	186	202	218	234	250
139	155	171	187	203	219	235	251
140	156	172	188	204	220	236	252
141	157	173	189	205	221	237	253
142	158	174	190	206	222	238	254
143	159	175	191	207	223	239	255

KO18-R

в системе MS DOS

	0	1	2	3	4	5	6	7	8	9	A	B	C	D	E	F
8	A	Б	В	Г	Д	Е	Ж	З	И	Й	К	Л	М	Н	О	П
9	Q	С	Т	У	Ф	Х	Ц	Ч	Ш	Щ	Ъ	Ы	Э	Ю	Я	
A	а	б	в	г	д	е	ж	з	и	й	к	л	м	н	о	п
B	q	с	т	у	ф	х	ц	ч	ш	щ	ъ	ы	э	ю	я	
C	1	2	3	4	5	6	7	8	9	0	.	,	+	-	*	/
D	00	01	02	03	04	05	06	07	08	09	0A	0B	0C	0D	0E	0F
E	10	11	12	13	14	15	16	17	18	19	1A	1B	1C	1D	1E	1F
F	20	21	22	23	24	25	26	27	28	29	2A	2B	2C	2D	2E	2F

CP-866

на компьютерах
фирмы Apple
(Макинтош и др.)

	.0	.1	.2	.3	.4	.5	.6	.7	.8	.9	.A	.B	.C	.D	.E	.F
В.	А	Б	В	Г	Д	Е	Ж	З	И	Й	К	Л	М	Н	О	П
У.	Р	С	Т	У	Ф	Х	Ц	Ч	Ш	Щ	Ъ	Ы	Ь	Э	Ю	Я
А.	†	°	ε	£	§	*	І	Ө	©	™	Ъ	ђ	≠	Г	ƒ	
В.	∞	≤	≤	≥	і	μ	Д	Ј	Є	Є	І	і	Љ	Њ	Ў	Ў
С.	j	S	≈	√	f	≈	Δ	»	...		Ъ	ђ	К	к	ѕ	
Д.	—	“	”	‘	’	+	„	Ў	У	Ц	и	№	Е	є		
Е.	а	б	в	г	д	е	ж	з	и	й	к	л	м	н	о	п
Ғ.	р	с	т	у	ф	х	ц	ч	ш	щ	ъ	ы	ь	э	ю	я

MacCyrillic

Windows-1251	KOI8-R
Привет, Вася!	oПХБЕР, бЮЯЪ!
рТЙЧЕФ, чБУС!	Привет, Вася!

Стандарт UNICODE

Для того чтобы в одном документе можно было использовать более 256 символов, в 1991 году был принят стандарт кодирования символов UNICODE, который включает знаки любых существующих (и даже некоторых мёртвых) языков, математические и музыкальные символы и др.

Windows

- **UTF-16**
- В ней все наиболее важные символы кодируются с помощью 16 бит (2 байт), а редко используемые — с помощью 4 байт.

Unix - Linux

- **UTF-8**
- В ней все символы, входящие в таблицу ASCII, кодируются с помощью 1 байта, а другие символы могут занимать от 2 до 4 байт. Если значительную часть текста составляют латинские буквы и цифры, такой подход позволяет значительно уменьшить объём файла по сравнению с UTF-16.

