

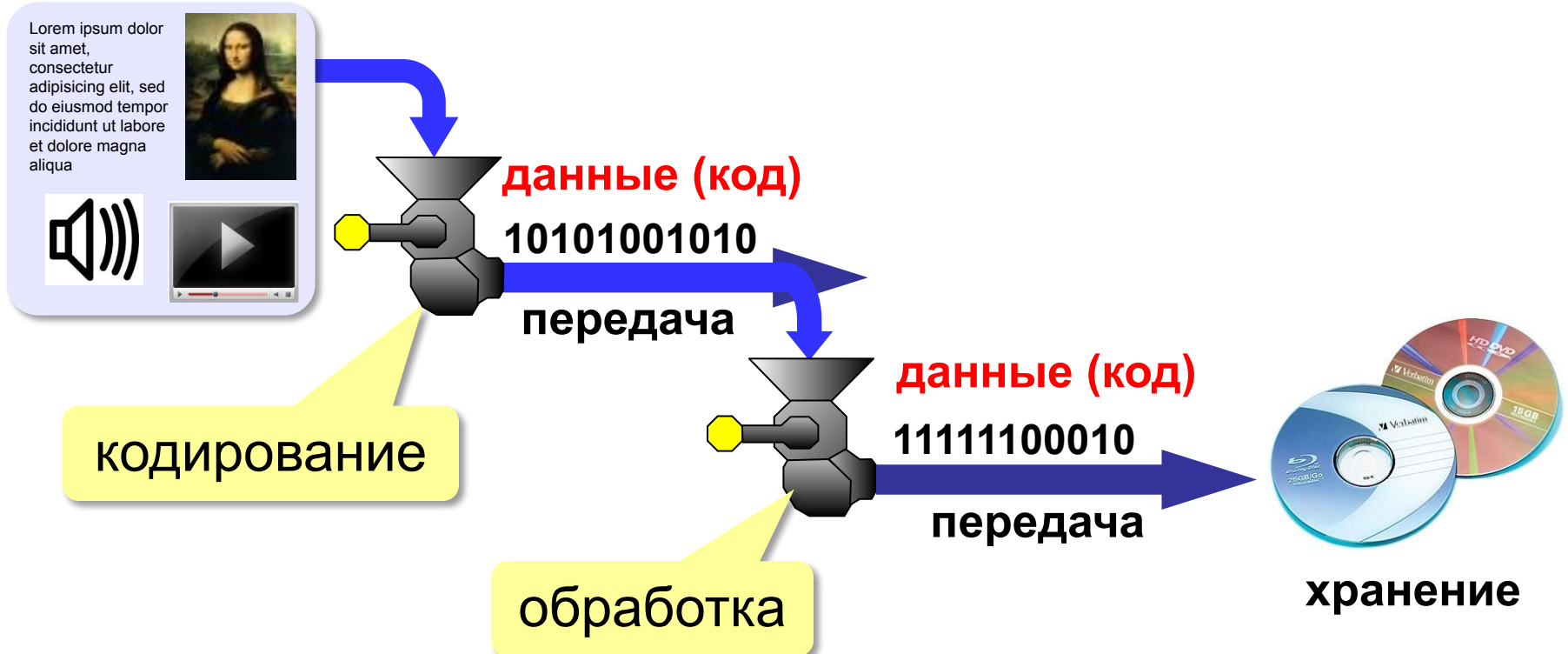
Кодирование информации

§ 13. Кодирование символов

Зачем кодировать информацию?

Кодирование — это представление информации в форме, удобной для её хранения, передачи и обработки.

В компьютерах используется двоичный код:

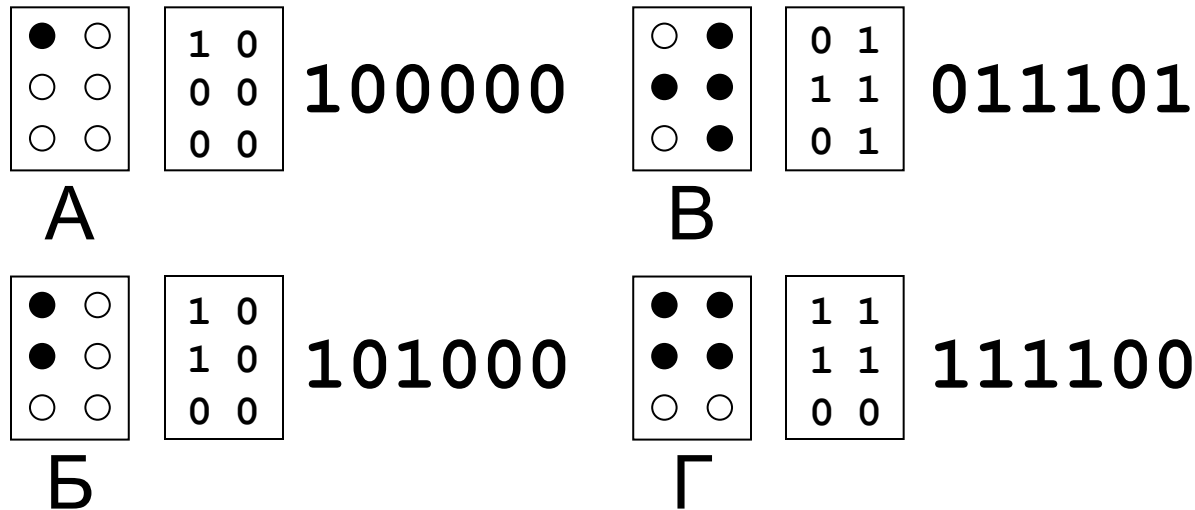


Кодирование информации

§ 13. Кодирование символов

Кодирование символов

Система Брайля:



Общий подход:

- нужно использовать N символов
- выберем число битов k на символ: $2^k \geq N$
- сопоставим каждому символу код — число от 0 до $2^k - 1$
- переведем коды в двоичную систему



Откуда формула?

Кодирование символов

Текстовый файл

- на экране (символы)
- в памяти — коды



1000001 ₂	1000010 ₂	1000011 ₂	1000100 ₂
65	66	67	68



В файле хранятся не изображения символов, а их числовые коды!

Файлы со шрифтами: *.fon, *.ttf, *.otf

Кодировка ASCII (7-битная)

ASCII = *American Standard Code for Information Interchange*

Коды 0-127:

0-31 **управляющие символы:**

7 – звонок, 10 – новая строка,

13 – возврат каретки, 27 – Esc.

32 пробел

знаки препинания: . , : ; ! ?

специальные знаки: + – * / () { } []

48-57 цифры **0..9**

65-90 заглавные латинские буквы **A-Z**

97-122 строчные латинские буквы **a-z**



Где русские буквы?

8-битные кодировки

Кодовые страницы (расширения ASCII):

0	1	127	128	254	255
		таблица ASCII		национальный алфавит	

Для русского языка:

CP-866 для *MS DOS*

CP-1251 для *Windows* (Интернет)

KOI8-R для *UNIX* (Интернет)

MacCyrillic для компьютеров *Apple*

Проблема:

Windows-1251

Привет, Вася!

рТЙЧЕФ,
чБУС!

KOI8-R

оПХБЕР,
бЮЯЪ!

Привет, Вася!

8-битные кодировки



- 1 байт на символ – файлы небольшого размера!
- просто обрабатывать в программах



- нельзя использовать символы разных кодовых страниц одновременно (русские и французские буквы, и т.п.)
- неясно, в какой кодировке текст (перебор вариантов!)
- для каждой кодировки нужен свой шрифт (изображения символов)

Стандарт UNICODE

1 112 064 знаков, используются около **100 000**

Windows: **UTF-16**

16 битов на распространённые символы,
32 бита на редко встречающиеся

Linux: **UTF-8**

8 битов на символ для ASCII,
от 16 до 48 бита на остальные



- совместимость с ASCII
- более экономична, чем UTF-16, если много символов ASCII



2010 г. – 50% сайтов использовали UTF-8!