|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Кодирование текстовой информации**   |  |  |  |  | | --- | --- | --- | --- | |  | http://school497.ru/download/u/02/img/text4.jpg |  | С точки зрения ЭВМ текст состоит из отдельных символов. К числу символов принадлежат не только буквы (заглавные или строчные, латинские или русские), но и цифры, знаки препинания, спецсимволы типа "=", "(", "&" и т.п. и даже (обратите особое внимание!) пробелы между словами. Да, не удивляйтесь: пустое место в тексте тоже должно иметь свое обозначение. |   **Вспомним некоторые известные нам факты:**  Множество символов, с помощью которых записывается текст, называется ***алфавитом***.  Число символов в алфавите – это его ***мощность***.  Формула определения количества информации: ***N = 2b***,  где N – мощность алфавита (количество символов),  b – количество бит (информационный вес символа).  В алфавит мощностью 256 символов можно поместить практически все необходимые символы. Такой алфавит называется ***достаточным.***  Т.к. 256 = 28, то вес 1 символа – 8 бит.  Единице измерения 8 бит присвоили название ***1 байт:***  1 байт = 8 бит.  ***Двоичный код каждого символа в компьютерном тексте занимает 1 байт памяти.***  **Каким же образом текстовая информация представлена в памяти компьютера?**   |  |  |  |  | | --- | --- | --- | --- | |  | http://school497.ru/download/u/02/img/text5.jpg |  | Тексты вводятся в память компьютера с помощью клавиатуры. На клавишах написаны привычные нам буквы, цифры, знаки препинания и другие символы. В оперативную память они попадают в двоичном коде. Это значит, что каждый символ представляется 8-разрядным двоичным кодом.  Кодирование заключается в том, что каждому символу ставится в соответствие уникальный десятичный код от 0 до 255 или соответствующий ему двоичный код от 00000000 до 11111111. Таким образом, человек различает символы по их начертанию, а компьютер - по их коду. |   Удобство побайтового кодирования символов очевидно, поскольку байт - наименьшая адресуемая часть памяти и, следовательно, процессор может обратиться к каждому символу отдельно, выполняя обработку текста. С другой стороны, 256 символов – это вполне достаточное количество для представления самой разнообразной символьной информации.  **Теперь возникает вопрос, какой именно восьмиразрядный двоичный код поставить в соответствие каждому символу.**  Понятно, что это дело условное, можно придумать множество способов кодировки.  Все символы компьютерного алфавита пронумерованы от 0 до 255. Каждому номеру соответствует восьмиразрядный двоичный код от 00000000 до 11111111. Этот код просто порядковый номер символа в двоичной системе счисления.  **Таблица, в которой всем символам компьютерного алфавита поставлены в соответствие порядковые номера, называется таблицей кодировки.**  Для разных типов ЭВМ используются различные таблицы кодировки.  Международным стандартом для ПК стала таблица ***ASCII***(читается аски) (Американский стандартный код для информационного обмена ***A****merican****s****tandard****c****ode for****i****nformation****i****nterchange*).  Таблица кодов ASCII делится на две части.  Международным стандартом является лишь первая половина таблицы, т.е. символы с номерами от ***0*** (00000000), до ***127*** (01111111).  **Структура таблицы кодировки ASCII**   |  |  |  | | --- | --- | --- | | **Порядковый номер** | **Код** | **Символ** | | **0 - 31** | **00000000 - 00011111** | Символы с номерами от 0 до 31 принято называть управляющими.         Их функция – управление процессом вывода текста на экран или печать, подача звукового сигнала, разметка текста и т.п. | | **32 - 127** | **00100000 - 01111111** | Стандартная часть таблицы (английский). Сюда входят строчные и прописные буквы латинского алфавита, десятичные цифры, знаки препинания, всевозможные скобки, коммерческие и другие символы.         Символ 32 - пробел, т.е. пустая позиция в тексте.         Все остальные отражаются определенными знаками. | | **128 - 255** | **10000000 - 11111111** | Альтернативная часть таблицы (русская).         Вторая половина кодовой таблицы ASCII, называемая кодовой страницей (128 кодов, начиная с 10000000 и кончая 11111111), может иметь различные варианты, каждый вариант имеет свой номер.         Кодовая страница в первую очередь используется для размещения национальных алфавитов, отличных от латинского. В русских национальных кодировках в этой части таблицы размещаются символы русского алфавита. |   **Первая половина таблицы кодов ASCII**   |  |  |  |  | | --- | --- | --- | --- | |  | http://school497.ru/download/u/02/img/asc2.gif |  |  |   Обращаю ваше внимание на то, что в таблице кодировки буквы (прописные и строчные) располагаются в алфавитном порядке, а цифры упорядочены по возрастанию значений. Такое соблюдение лексикографического порядка в расположении символов называется принципом последовательного кодирования алфавита.  Для букв русского алфавита также соблюдается принцип последовательного кодирования. |
| **Вторая половина таблицы кодов ASCII**  **http://school497.ru/download/u/02/img/asc1.gif**  К сожалению, в настоящее время существуют пять различных кодировок кириллицы (КОИ8-Р, Windows. MS-DOS, Macintosh и ISO). Из-за этого часто возникают проблемы с переносом русского текста с одного компьютера на другой, из одной программной системы в другую.  Хронологически одним из первых стандартов кодирования русских букв на компьютерах был КОИ8 ("Код обмена информацией, 8-битный"). Эта кодировка применялась еще в 70-ые годы на компьютерах серии ЕС ЭВМ, а с середины 80-х стала использоваться в первых русифицированных версиях операционной системы UNIX.  От начала 90-х годов, времени господства операционной системы MS DOS, остается кодировка CP866 ("CP" означает "Code Page", "кодовая страница").  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.  Наиболее распространенной в настоящее время является кодировка Microsoft Windows, обозначаемая сокращением CP1251.  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.  Попробуем с помощью таблицы ASCII представить, как будут выглядеть слова в памяти компьютера.  Внутреннее представление слов в памяти компьютера   |  |  | | --- | --- | | **Слова** | **Память** | | **file** | **01100110**  **01101001**  **01101100**  **01100101** | | **disk** | **01100100**  **01101001**  **01110011**  **01101011** |   Иногда бывает так, что текст, состоящий из букв русского алфавита, полученный с другого компьютера, невозможно прочитать - на экране монитора видна какая-то "абракадабра". Это происходит оттого, что на компьютерах применяется разная кодировка символов русского языка. |