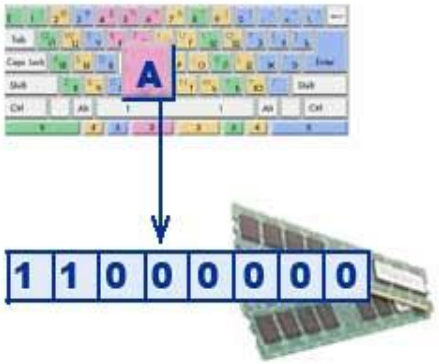
Тема урока: **Кодирование символьной информации**



Для обработки текстовой информации на компьютере необходимо представить ее в двоичной знаковой системе.

Для кодирования каждого знака требуется количество информации, равное 8 битам, т. е. длина двоичного кода знака составляет восемь двоичных знаков.

Каждому знаку необходимо поставить в соответствие уникальный двоичный код из интервала от 00000000 до 11111111 (в десятичном коде от 0 до 255

Человек различает знаки по их начертанию, а компьютер - по их двоичным кодам. При вводе в компьютер текстовой информации происходит ее двоичное кодирование, изображение знака преобразуется в его двоичный код.

Пользователь нажимает на клавиатуре клавишу со знаком, и в компьютер поступает определенная последовательность из восьми электрических импульсов (двоичный код знака).

Код знака хранится в оперативной памяти компьютера, где занимает одну ячейку.

В процессе вывода знака на экран компьютера производится обратное перекодирование, т. е. преобразование двоичного кода знака в его изображение.

Различные кодировки знаков. Присваивание знаку конкретного двоичного кода - это вопрос соглашения, которое фиксируется в кодовой таблице. В существующих кодовых таблицах первые 33 кода (десятичные коды с 0 по 32) соответствуют не знакам, а операциям (перевод строки, ввод пробела и т. д.).

ASCII — базовая кодировка текста для латиницы

Первоначально для персональных компьютеров был взят за основу так называемый **ASCII-код разработанный и стандартизированный в США в 1963 г.** ***(American Standard Code for Information Interchange).***

Этот код содержит 7 бит информации и в нем можно представить 128 различных комбинаций для кодирования символов. Этого вполне достаточно для того, чтобы закодировать заглавные и строчные буквы латинского алфавита, цифры, знаки препинания и ряд специальных и управляющих символов.
Информация в Internet до сих пор передается в 7 битном коде. Затем к 7-ми битному добавили еще один – восьмой бит, что позволило закодировать еще 128 символов (всего 256), которые предназначались для символов псевдографики и национальных шрифтов, которые опять-таки могут иметь в своей основе латиницу, кириллицу (напр. Болгарский, Русский) или другое (напр. Греческое) начертание символов – **расширенный ASCII-код**.

Присвоение символу конкретного двоичного кода – это вопрос соглашения, которое фиксируется в кодовой таблице.

Первые 32 кода (с 0 до 31) 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.

Коды от 32 до 127. 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.

Коды с 128 по 255 являются национальными, т.е. в национальных кодировках одному и тому же коду соответствуют различные символы. 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.

К сожалению, в настоящее время существует более 6 различных кодовых таблиц для русских букв, поэтому тексты созданные в одной кодировке, не будут правильно отображаться в другой.

1. КОИ-8

Хронологически одним из первых стандартов кодирования русских букв на компьютерах был код **КОИ – 8 («Код обмена информационный – 8 битный»).**

KOI-8 — это восьмибитная кодовая страница, совместимая с **ASCII**. Разработана для кодирования букв кириллических алфавитов.

Была широко распространена как основная русская кодировка в Unix-совместимых ОС и в электронной почте, однако к концу 2010 г. с распространением Юникода, постепенно вышла из употребления.

2. CP 866

аботы в среде операционной системы
OS используется «альтернативная»
овка, в терминологии фирмы Microsoft –
овка **CP 866**. **(Code Page 866).**

ая кодировка является расширением
овки ASCII

3. Windows-1251

lows-1251 (CP 1251) *(«CP» означает «Code Page»).* **Все lows – приложения, работающие с русским языком, ерживают эту кодировку.**

ставляет собой набор символов и кодов, является артной 8-битной кодировкой для русских й Microsoft Windows до 10-й версии.

ошлом пользовалась довольно большой популярностью. Была создана на базе кодировок, использовавшихся в ранних «самопальных» русификаторах Windows в 1990—1991 гг. совместно представителями «Параграфа», «Диалога» и российского отделения Microsoft.

В современных приложениях отдается предпочтение Юникоду (UTF-8). На 7 июля 2016 лишь на 1.8% всех веб-страниц используется **Windows-1251.**

4. MacCyrillic (MAC)

Кодировка **MacCyrillic** используется только на компьютерах "Макинтош".

Благодаря отсутствию псевдографики и "верхних" управляющих символов эта кодировка включает довольно много полезных символов; кроме того, присутствуют все дополнительные буквы, необходимые для записи украинского, белорусского, македонского и сербского языков.

5. ISO 8859-5

Международная организация по стандартизации (International Standards Organization, ISO) утвердила в качестве стандарта для русского языка еще одну кодировку под названием **ISO 8859 – 5.**

ISO 8859-5 — 8-битная кодовая страница из семейства кодовых страниц стандарта ISO-8859 для представления кириллицы.

Имеются буквы многих языков, использующих кириллицу, однако в целом ISO 8859-5 — не очень удобная кодировка, поскольку в ней отсутствуют многие нужные символы, такие как тире (—), кавычки-ёлочки («»), градус (°) и др.

Порядок символов этой кодовой страницы использовался при размещении букв кириллицы в наборе символов **Unicode** (со сдвигом вверх на 864 позиции).

Применение

ISO 8859-5 широко применяется в Сербии и иногда в Болгарии на юниксоподобных системах.

6. UNICODE

Юникод (*Unicode*) — стандарт кодирования символов, позволяющий представить знаки почти всех письменных языков.

Стандарт предложен в 1991 году некоммерческой организацией «Консорциум Юникода» (*Unicode Consortium, Unicode Inc.*).

Стандарт отводит на каждый символ не один байт, а два, и поэтому с его помощью можно закодировать не 256 символов, **2^{16}=65 536 различных символов.** Эту кодировку поддерживает платформа **Microsoft Windows и Microsoft Office** *(вставка символов)*.

Применение этого стандарта позволяет закодировать очень большое число символов из разных письменностей: в документах Unicode могут соседствовать китайские иероглифы, математические символы, буквы греческого алфавита, латиницы и кириллицы, при этом становится ненужным переключение кодовых страниц.

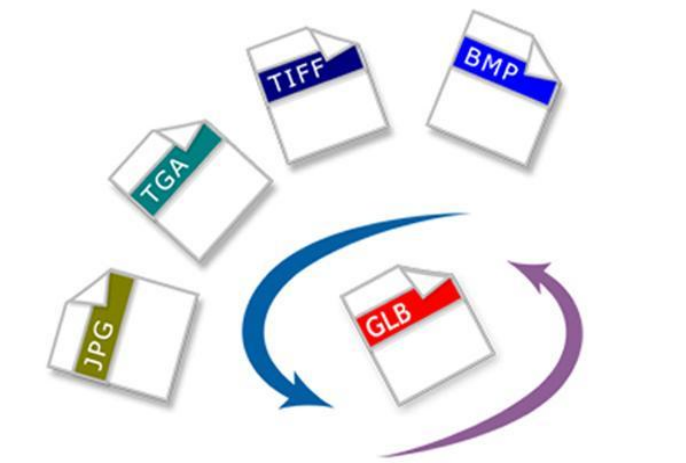
Стандарт состоит из двух основных разделов:

1. ***универсальный набор символов (UCS, universal character set)***
2. ***семейство кодировок (UTF, Unicode transformation format).***

Универсальный набор символов задаёт однозначное соответствие символов кодам — элементам кодового пространства, представляющим неотрицательные целые числа.

Семейство кодировок определяет машинное представление последовательности кодов UCS.

РАЗЛИЧНЫЕ ФОРМАТЫ ТЕКСТОВЫХ ФАЙЛОВ



Формат файла или тип файла определяет способ кодирования информации в файле (перевода в двоичный код).

В текстовом файле помимо кодировки символов кодируются операции, обеспечивающие форматирование текста.

В различных текстовых редакторах символы форматирования кодируются по разному, документы различных текстовых редакторов имеют разный формат (тип).

ОСНОВНЫЕ ФОРМАТЫ ТЕКСТОВЫХ ФАЙЛОВ

1.TXT *(Text Only)*. Это старейший текстовый формат, аналоги современного блокнота были еще на первых ПК. Является наиболее универсальным. Документы txt открываются текстовыми редакторами, работающими в любой операционной системе. Формат очень простой и не содержит ничего, кроме текста. Форматирование не поддерживается — сохраняются только абзацы, отступ и заглавные буквы. Поэтому файлы-txt отличаются маленькими размерами. Формат устойчив к повреждениям. При повреждении части файла можно восстановить или обработать остальную часть документа.

2. DOC *(Document)* Формат редактора **MS Word 97-2003**. Использует кодировку **UNICODE**.

Сегодня формат **doc** предоставляет огромные возможности по обработке текста и вставке в документ различных изображений, диаграмм, таблиц, ссылок. Может включать в себя сценарии и макросы. *Но формат является закрытым, многие документы в этом формате корректно отображаются только в самой программе* ***MS Word.***

3. RTF *(Rich Text Format)*. Специально разработан программистами компаний **Microsoft и Adobe** для обмена файлами между пользователями.

Может быть открыт и обработан на любой платформе. Поддерживается многими приложениями.

Документы **rtf** поддерживает сложное форматирование. Помимо текста может содержать различные рисунки, таблицы, вставки и сноски. В нем могут использоваться несколько видов шрифтов. Формат устойчив к повреждению файлов. Так как в **rtf** не используются макросы, он считается более безопасным чем формат **doc.**

4. HTM, HTML (Hyper Text Markup Language). Формат разметки Web-страниц. Содержит управляющие коды (тэги) языка разметки гипертекста.

5. DOCX Формат редактора Microsoft Office 2007 и следующих версиях.

Впервые был применен в MS Word 2007. Его главное отличие от формата doc — ***использование zip-компрессии*** для уменьшения объема файла.

Представляет собой архив с данными, содержащий помимо текста в формате XML, изображения, стили текста, форматирование и другие данные.

Причем текстовые файлы и графика хранятся в отдельных документах.

Чтобы увидеть содержимое docx-файла можно изменить его расширение на **zip** и открыть в любом архиваторе.

Чтобы открыть документ-docx в ранних версиях Word, необходимо скачать и установить «Пакет обеспечения совместимости Microsoft Office для форматов файлов Word, Excel и PowerPoint»

6. ODT/ODF (Open Document Format)– формат для текстового редактора **Write** пакета свободный офис (**OpenOffice)**

Это открытый формат, который может использоваться без ограничений и является альтернативой форматам Microsoft.

7. CHM – формат файла, который используется для хранения нескольких файлов формата **html.**

Создан для замены формата справочной системы. Может иметь ссылки, по которым возможен переход на другую страницу.