Кодирование символов Unicode

Любые числа (в определенных пределах) в памяти компьютера кодируются числами двоичной системы счисления. Для этого существуют простые и понятные правила перевода. Однако на сегодняшний день компьютер используется куда шире, чем в роли исполнителя трудоемких вычислений. Например, в памяти ЭВМ хранятся текстовая и мультимедийная информация. Поэтому возникает первый вопрос:

Как в памяти компьютера хранятся символы (буквы)?

Каждая буква принадлежит определенному алфавиту, в котором символы следуют друг за другом и, следовательно, могут быть пронумерованы последовательными целыми числами. Каждой букве можно сопоставить целое положительное число и назвать его кодом символа. Именно этот код будет храниться в памяти компьютера, а при выводе на экран или бумагу «преобразовываться» в соответствующий ему символ. Чтобы отличить представление чисел от представления символов в памяти компьютера, приходится также хранить информацию о том, какие именно данные закодированы в конкретной области памяти.

Соответствие букв определенного алфавита с числами-кодами формирует так называемую таблицу кодирования. Другими словами, каждый символ конкретного алфавита имеет свой числовой код в соответствии с определенной таблицей кодирования.

Однако алфавитов в мире очень много (английский, русский, китайский и др.). Поэтому следующий вопрос:

Как закодировать все используемые на компьютере алфавиты?

Для ответа на этот вопрос пойдем историческим путем.

В 60-х годах XX века в американском национальном институте стандартизации (ANSI) была разработана таблица кодирования символов, которая впоследствии была использована во всех операционных системах. Эта таблица называется
ASCII
(
American
Standard
Code
for
Information
Interchange
– американский стандартный код для обмена информацией). Чуть позже появилась расширенная версия
ASCII
.

В соответствие с таблицей кодирования ASCII для представления одного символа выделяется 1 байт (8 бит). Набор из 8 ячеек может принять 2
8
= 256 различных значений. Первые 128 значений (от 0 до 127) постоянны и формируют так называемую основную часть таблицы, куда входят десятичные цифры, буквы латинского алфавита (заглавные и строчные), знаки препинания (точка, запятая, скобки и др.), а также пробел и различные служебные символы (табуляция, перевод строки и др.). Значения от 128 до 255 формируют дополнительную часть таблицы, где принято кодировать символы национальных алфавитов.

Поскольку национальных алфавитов огромное множество, то расширенные ASCII-табли цы существуют во множестве вариантов. Даже для русского языка существуют несколько таблиц кодирования (распространены
Windows
-1251 и
Koi
8-
r
). Все это создает дополнительные трудности. Например, мы отправляем письмо, написанное в одной кодировке, а получатель пытается прочитать ее в другой. В

результате видит кракозябры. Поэтому читающему требуется применить для текста другую таблицу кодирования.

Есть и другая проблема. В алфавитах некоторых языков слишком много символов и они не помещаются в отведенные им позиции с 128 до 255 однобайтовой кодировки.

Третья проблема - что делать, если в тексте используется несколько языков (например, русский, английский и французский)? Нельзя же использовать две таблицы сразу …

Чтобы решить эти проблемы одним разом была разработана кодировка Unicode.

Стандарт кодирования символов Unicode

Для решения вышеизложенных проблем в начале 90-х был разработан стандарт кодирования символов, получивший название Unicode. Данный стандарт позволяет использовать в тексте почти любые языки и символы.

В Unicode для кодирования символов предоставляется 31 бит (4 байта за вычетом одного бита). Количество возможных комбинаций дает запредельное число: 231 = 2 147 483 684 (т.е. более двух миллиардов). Поэтому Unicode описывает алфавиты всех известных языков, даже «мертвых» и выдуманных, включает многие математические и иные специальные символы. Однако информационная емкость 31-битового Unicode все равно остается слишком большой. Поэтому чаще используется сокращенная 16-битовая версия (216 = 65 536 значений), где кодируются все современные алфавиты.

В Unicode первые 128 кодов совпадают с таблицей ASCII.