

Кодирование

Определения

Кодирование информации - процесс преобразования сигнала из формы удобной для непосредственного использования информации в форму удобную для *передачи, хранения* или автоматической *обработки*.

Кодирование (как действие?) - это операция отождествления символов или групп символов одного кода с символами и группами символов другого кода.

Для кодирования нам необходима *специальная таблица*.

Необходимость кодирования возникает прежде всего из потребности приспособить форму сообщения к данному каналу связи или какому-либо другому устройству, предназначенному для преобразования или хранения данных.

Цель кодирования

Цель кодирования (в теории информации) состоит в уменьшении избыточности сообщений и влияния помех, искажающих сообщение при передачи по каналу связи.

- Избыточность
- Помехи

Передача информации - физический процесс, по средством которого осуществляется перемещение знаков (сведений, способных предоставлять информацию) в пространстве или осуществляется физический доступ субъектов к знакам.

Кодирование символов - процесс присвоения чисел графическим символам, что позволяет хранить, передавать и преобразовывать их с помощью компьютеров.

Набор символов (*character set*) - таблица, задающая кодировку, конечного множества символа алфавита (обычно элементов текста: букв, цифр, знаков препинания).

Такая таблица сопоставляет каждому символу последовательность длиной в 1 или несколько символов другого алфавита (точки и тире в азбуке Морзе; сигнальные флаги на флоте; нули единицы (биты) в компьютерах).

Символы в компьютере обычно кодируются одним или несколькими байтами

ASCII (*American Standartized Code for International Interchange*) - стандарт кодирования букв латинского алфавита, цифр, некоторых специальных знаков и управляющих символов, принятый в 1963 году американской ассоциацией стандартов как основной способ представления данных в ЭВМ.

ASCII - 7 битный код, содержит $2^7 = 128$ кодовых позиций, в которых размещены следующие символы:

1. десятичные цифры

2. латинские буквы
3. знаки препинания
4. орфографические знаки
5. математические символы
6. управляющие символы

Важная особенность стандарта - не используется *переключения регистра* (верхний и нижние регистры - разные символы).

В стандарте ASCII есть группа управляющих символов, которые в будущем стали ненужными.

Следующая кодировка - **UTF-8** (доминирующая на сегодняшний день). **UTF-8** (*Unicode Transformation Format*) - распространённый стандарт кодирования символов, позволяющий более компактно хранить и передавать символы Unicode, используя переменное количество байт и обеспечивающий полную обратную совместимость с устаревшим ASCII.

Внимание

Unicode - набор символов

UTF-8 - кодировка

Кодировка **не означает** набор символов

В связи с появлением новых кодировок появлялись проблемы:

- *проблема неправильной раскодировки* - вызывало появления в документе символов иностранных слов, не предполагавшимся в документе, появление **псевдосимволов? (кракозябры)**
- проблема ограниченности набора символов
- проблема преобразования одной кодировки в другую
- проблема дублирования шрифтов

Кодовое пространство Unicode разбито на 17 плоскостей по 2^{16} символов.

Нулевая плоскость называется *базовой* и содержит символы наиболее употребительных письменностей (остальные плоскости - *дополнительные*).

Принципы Unicode:

1. *Гарантии стабильности* (как только символ появился в кодировке, он не сдвинется и не исчезнет. Каждая новая версия Unicode будет содержать символы прошлых версий).

2. Если символ окажется "*плохим*", его запрещают.

3. *Простой тест*

Такой высокой целью, как универсальность, Unicode добивается *динамической сборки символов*. Unicode кодирует простой текст без оформления. Считается, что простой текст должен хранить достаточно данных, чтобы читаемо отобразить его и ничего более.

4. *Универсальность* (Unicode разработан для людей разных языков, профессий, для современных и исторических текстов)

За пределами Unicode лежат:

1. письменности, про которые мало что известно, чтобы надёжно закодировать символы.

2. письменности, чьи пользователи не пришли к de facto стандарту.

3. не текстовые письменности (пиктография)

5. *Унификация*. Unicode старается не дублировать символы. (сходные символы разных письменностей кодируются по-разному!)

Консорциум Unicode не создаёт нового, а констатирует сложившийся порядок вещей. Символ российского рубля прошел согласование его включения в Unicode. Причем он много лет до

этого использовался, но когда он получил официальное утверждение, его включили в Unicode.