Распространенные русские кодировки

Из русской «Википедии»

Содержание

1 IBM866, или «альтернативная кодировка»

«Альтернативная кодировка» — основанная на CP437 кодовая страница, где все специфические европейские символы во второй половине заменены на кириллицу, оставляя псевдографические символы нетронутыми. Следовательно, это не портит вид программ, использующих для работы текстовые окна, а также обеспечивает использование в них символов кириллицы.

Исторически существовало много вариантов альтернативной кодировки, но все различия касаются только области 0xF0–0xFF (240–255). Окончательным стандартом стала кодировка IBM CP866, поддержка которой была добавлена в MS-DOS версии 6.22 (до этого использовались всевозможные «самопальные» русификаторы). Альтернативная кодировка всё ещё жива и чрезвычайно популярна в среде DOS и OS/2. Кроме того, в этой кодировке записываются имена в файловой системе FAT (и короткие имена в VFAT). CP866 до сих пор используется в консоли русифицированных систем семейства Windows NT.

2 Кодировка KOI8-R

КОИ-8 (код обмена информацией, 8 битов), KOI8 — восьмибитовая ASCII-совместимая кодовая страница, разработанная для кодирования букв кириллических алфавитов.

Существует также семибитовая версия кодировки, не полностью совместимая с ASCII — КОИ-7. КОИ-7 и КОИ-8 описаны в ГОСТ 19768-74 (сейчас недействителен).

Разработчики КОИ-8 поместили символы русского алфавита в верхней части кодовой таблицы таким образом, что позиции кириллических символов соответствуют их фонетическим аналогам в английском алфавите в нижней части таблицы. Это означает, что если в тексте, написанном в КОИ-8, убирать восьмой бит каждого символа, то получается «читабельный» текст, хотя он и написан латинскими символами. Например, слова «Русский Текст» превратились бы в «rUSSKIJ tEKST». Как побочное следствие, символы кириллицы оказались расположены не в алфавитном порядке.

Существует несколько вариантов кодировки КОИ-8 для различных кириллических алфавитов. Русский алфавит описывается в кодировке KOI8-R, украинский — в KOI8-U.

KOI8-R стал фактически стандартом для русской кириллицы в юниксподобных операционных системах и электронной почте.

Андрей Чернов создал документ RFC 1489 («Registration of a Cyrillic Character Set»), который, однако, не относится к категории RFC-стандартов. Существует RFC 2319 на KOI8-U.

По набору символов KOI8-R соответствует одному из вариантов альтернативной кодировки. Стандарт RFC 1489 также предписывает наличие графических символов «рамок» (псевдографики), однако это требование выполняется довольно редко.

3 Кодировка windows-1251

Windows-1251 — набор символов и кодировка, являющаяся стандартной 8-битной кодировкой для всех русских версий Microsoft Windows. Пользуется 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. Первоначальный вариант кодировки сильно отличался от представленного в таблице на Википедии (в частности, там было значительное число «белых пятен»).

Windows-1251 выгодно отличается от других 8-битных кириллических кодировок (таких как CP866, KOI8-R и ISO-8859-5) наличием практически всех символов, использующихся в русской типографике для обычного текста (отсутствует только значок ударения); она также содержит все символы для близких к русскому языку языков: украинского, белорусского, сербского и болгарского.

Если к кириллическому тексту в кодировке Windows-1251 20 раз подряд применить перекодирование KOI8-R ⟶ Windows-1251, в итоге будет получен исходный текст.

Имеет два недостатка:

- строчная буква «я» имеет код 0xFF (255 в десятичной системе). Она является «виновницей» ряда неожиданных проблем в программах без

поддержки чистого 8-го бита, а также (гораздо более частый случай) использующих этот код как служебный (в CP437 он обозначает «неразрывный пробел», в Windows-1252 — ÿ, оба варианта практически не используются; число же -1, в байтовом представлении аналогичное 255, часто используется в программировании как «пустое значение»);

- отсутствуют символы псевдографики, имеющиеся в CP866 и KOI8 (хотя для самих Windows, для которых она предназначена, в них не было нужды, это делало несовместимость двух использовавшихся в них кодировок заметнее).

4 Кодировка ISO-8859-5

ISO 8859-5 — 8-битная кодовая страница из серии ISO-8859 для представления кириллицы. В России почти не употребляется.

ISO 8859-5 была создана на базе «основной кодировки» (все русские буквы сохранили своё расположение, за исключением заглавной Ё).

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. Нет также буквы «CYRILLIC LETTER GHE WITH UPTURN», используемой иногда в украинской письменности.

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