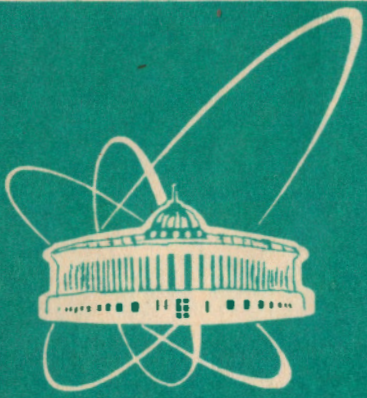

сообщения
объединенного
института
ядерных
исследований
дубна

Р11-93-23

В.В.Галактионов

РАБОТА С РУССКОЯЗЫЧНЫМИ ТЕКСТАМИ НА ЭВМ ТИПА VAX

1993

Решение проблемы работы с русскозычными и смешанными текстами на любой ЭВМ зависит от:

- выбора двоичного представления (кодировки) русских символов,
- наличия терминального и другого (в частности, принтера) оборудования, способного воспроизводить символы,
- наличия системного программного обеспечения для обработки текстов с выбранной кодировкой русских символов.

При применении зарубежной вычислительной техники в ряде случаев работа с русскими текстами представляет определенные, иногда значительные, трудности. В большинстве случаев латинские символы на этих ЭВМ представляются в семиразрядной кодировке ASCII, и системное обеспечение обработки текстов (трансляторы, редакторы) ориентировано на это представление текстов. К тому же, терминальное оборудование, без специальных заказов, не предусматривает возможности ввода и воспроизведения русских символов (кириллицы). При этом значительные трудности также представляет решение проблемы обработки смешанных текстов (больших и малых латинских и русских символов).

В некоторых случаях на практике применяется паллиативное решение: при возможности какого-либо способа воспроизведения русских символов, чтобы хоть как-то обеспечить минимальную обработку русскоязычных текстов, проводится интерпретация семиразрядных кодов части латинских символов в качестве русских. При этом возможны варианты представления текстов:

- только русских больших и малых символов - при интерпретации всех кодов латинских символов в качестве русских,
- смешанных текстов больших латинских и русских символов - при интерпретации кодов малых латинских символов в качестве больших русских.

При этих представлениях символов возможна обработка таких текстов стандартными средствами операционных систем.

В данной работе предлагается вариант решения указанной задачи, реализованный на ЭВМ типа VAX в операционных системах VMS и UNIX. В общем случае программная реализация данной проблемы не

является ориентированной на определенный тип ЭВМ или ОС. Скорее всего - это терминалоориентированная работа, поскольку для зарубежных терминалов ввод и воспроизведение русских текстов - весьма проблематичны и, в общем случае, - неразрешимы. В каждом конкретном случае проблема знакогенерации решается индивидуально, несмотря на наметившиеся тенденции функциональной стандартизации и унификации терминальных устройств. Примером тому может служить распространение разного типа эмуляций (микропрограммных либо программируемых) типов VT100 и VT200 для самого широкого круга терминального оборудования, вплоть до персональных ЭВМ. В этом плане предлагаемая работа может служить также для изучения проблемы в качестве методического пособия.

1. Выбор способа кодирования русских символов

В предлагаемой разработке применяются два наиболее распространенных вида представления в восьмиразрядном двоичном коде русских символов - кодировка КОИ-8 и так называемая альтернативная кодировка, применяемая в операционных системах MS DOS для персональных ЭВМ, совместимых с IBM PC. Для смешанных текстов кодировка латинских символов совпадает с семиразрядным кодом ASCII. Ниже в таблицах представлены коды русских символов в обеих кодировках.

Кодировка КОИ-8

	00	01	02	03	04	05	06	07	08	09	0a	0b	0c	0d	0e	0f
80																
90																
a0																
b0																
c0	ю	а	б	ц	д	е	ф	г	х	и	й	к	л	м	н	о
d0	п	я	р	с	т	у	ж	в	ь	ы	з	ш	э	щ	ч	
e0	Ю	А	Б	Ц	Д	Е	Ф	Г	Х	И	Й	К	Л	М	Н	О
f0	П	Я	Р	С	Т	У	Ж	В	Ь	Ы	З	Ш	Э	Щ	Ч	

Альтернативная кодировка

	00	01	02	03	04	05	06	07	08	09	0a	0b	0c	0d	0e	0f
80	А	Б	В	Г	Д	Е	Ж	З	И	Й	К	Л	М	Н	О	П
90	Р	С	Т	У	Ф	Х	Ц	Ч	Ш	Щ	Ъ	Ы	Ь	Э	Ю	Я
a0	а	б	в	г	д	е	ж	з	и	й	к	л	м	н	о	п
b0																
c0																
d0																
e0	р	с	т	у	ф	х	ц	ч	ш	щ	ъ	ы	ь	э	ю	я
f0																

2. Знакогенерация

Для современных терминалов существует понятие "Soft Charachter Sets" - система для программируемого задания произвольных символьных образов "fonts". Идея знакогенерации заключается в задании кодов терминальных команд в виде ESC-последовательности двоичных кодов, формирующих образ символа на экране терминала из матричного представления символа. В этом разделе будет описан способ программного формирования символов для терминала CIT224+ фирмы C. ITOH Electronics для режима эмуляциии VT100-VT200.

2.1. Формирование и загрузка образов символов

Для терминалов семейства CIT224 существуют два способа задания и загрузки образов символов: метод CIET и DEC-compatible метод. Выбранный в данной разработке метод CIET предоставляет больше свободы действий при формировании образов символов на основе матрицы 10 X 16. При работе терминала в режиме 25 строк на экран под символ отводится матрица 7 X 15, при выключенной 25 строке - матрица 7 X 16. Каждый элемент матрицы соответствует одному пикселю на экране. Единичному значению элемента соответствует светящийся пиксель. Семиэлементная строка матрицы (с добавлением 8-го с нулевым значением) интерпретируется как шестнадцатеричное числовое значение байта. Для полной матрицы символа таких чисел будет 16. Это и есть основа для числового описания образа символа.

Для загрузки образа символа формируется ESC-последовательность:

DCS Pfn;Pcn;Pe w <name> <data> ST,

где:

DCS (device control string) - стартовый код командной строки,
Pfn - индекс буфера образов символов,
Pcn - начальный индекс символа (в десятичном представлении семиразрядный код символа),
Pe - флаг стирания матрицы символа,
w - символ-ограничитель,
<name> - трехсимвольное имя системы символьных образов,
<data> - числовое задание элементов матрицы,
ST - код-ограничитель для строки DCS.

Поле <data> может содержать целый блок последовательных числовых образов символов. При выполнении операции вывода созданной таким образом ESC-последовательности в памяти терминала будет сформирован "Soft character set" с указанным именем <name>.

Для дальнейших действий по актуализации (вводу в действие) загруженных образов выполняется операция закрепления их за одним из разделов памяти терминала - G0, G1, G2 или G3. Выбор раздела диктуется различного рода соображениями: числом блоков образов, удобством переключений генератора символов, использованием собственных графических средств терминала в программе пользователя и многими другими обстоятельствами. В предлагаемой работе был выбран раздел G3, что явилось вполне достаточным и приемлемым для последующих работ автора по программному обеспечению обработки русскоязычных текстов в системе GALAXY /1/ и текстового редактора VEGA /2/.

Эта процедура выполняется также выдачей на терминал ESC-последовательности закрепления именованного загруженного блока образов <name> за выбранным разделом памяти знакогенератора.

2.2. Ввод в действие (актуализация) Soft character set

Для аппаратного знакогенератора терминала используются две области задания кодов символов GL (grafic left) и GR (grafic right). По умолчанию (by factory default) область GL соответствует латинским символам в семиразрядной кодировке ASCII, а область GR - восьмиразрядным (начиная с шестнадцатеричного кода 80) кодам специальных графических символов.

Следующим шагом задания программируемых образов является "ввод в действие" выбранного раздела G0, G1, G2 или G3 или же "приписка" его к области GL или GR для определения типа кодировки символов (семи- или восьмиразрядной).

Эта процедура также выполняется выдачей на терминал ESC-последовательности приписки.

В программах обработки смешанных текстов применяется приписка к восьмиразрядной области GR, поскольку область GL остается для использования семиразрядных символов ASCII.

В программах, где выполняется интерпретация семиразрядных кодов латинских символов (части или полностью) в качестве русских, приписка раздела выполняется к области семиразрядных кодов GL. Итак, для программного задания знакогенерации необходимо выполнить последовательность процедур:

- сформировать 7 X 16 элементную матрицу для каждого символа,
- дать числовое описание каждой матрицы,
- выполнить загрузку блоков образов символов в память терминала,
- выполнить операцию закрепления блока образов (fonts) за одним из разделов Gi,
- произвести актуализацию запрограммированных символов.

В большинстве случаев этого достаточно для воспроизведения на экране терминала начертаний символов при выдаче соответствующих кодов из ЭВМ. Например, просматривать тексты файлов либо MAIL-сообщения простыми средствами, типа команды Type для файлов

Для ввода же с клавиатуры русских букв этого недостаточно, поскольку клавиатурная кодировка вводимых символов - семиразрядная и не может перекрыть весь диапазон кодов для смешанных текстов. Эту задачу выполняет набор программных средств в виде подпрограмм-функций.

3. Программные средства обеспечения работы с русскоязычными текстами

Для каждого вида работ с русскими символами разработан свой программный набор.

3.1. Формирование образов символов

Для автоматизации процесса формирования числового описания матриц с образами символов разработаны две программы CR_SINGLE_FONT.EXE (для одиночного задания символа) и CR_GROUP_FONT.EXE (для блочного задания группы символов). Исходным материалом для программ является текстовый файл FONT.IMG с простым представлением матрицы (или множества матриц) символов как набора точек из 16 строк по 8 в строке. Значащая часть матрицы обозначена символом, отличным от точки.

Результат выполнения программ - текстовый файл HEX_FONT.H с полностью подготовленными ESC-последовательностями для загрузки образов символов в память терминала.

Таких файлов, как исходных, так и результирующих, готовится несколько - для каждой группы символьных образов, и хранятся они под разными именами. Например, исходные файлы KOI8_FONT.IMG (для русских символов в кодировке КОИ-8), MIX_BIG_FONT.IMG (для режима интерпретации кодов малых латинских символов для больших русских), RUS_FONT.IMG (для полной интерпретации всех кодов латинских символов для больших и малых русских), результирующие файлы KOI8_FONT.H, MIX_BIG_FONT.H и RUS_FONT.H с соответствующими ESC-последовательностями.

Процедуры формирования образов - одноразового использования, выполняются они в командном режиме (типа DCL); результат их работы применяется многократно: после выключения/включения терминалов, при загрузке новых образов - в динамическом режиме в процессе счета задачи.

3.2. Процедуры загрузки и актуализации образов символов

На основе подготовленных вышеуказанным способом ESC-последовательностей процедуры загрузки образов в память терминалов выполняют подпрограммы-функции:

set_koi8_font(), set_mix_font(), set_rus_font()

для соответствующих режимов интерпретации русских символов.

Функция set_font(), выдавая терминалу соответствующую командную ESC-последовательность, закрепляет загруженный блок образов за разделом памяти знакогенератора G0 и приписывает этот

раздел к области семиразрядных кодов GL. (Раздел G0 по умолчанию соответствует области GL и не требуется для этого больше специальных команд). Этот режим применяется при различного рода интерпретациях семиразрядных кодов латинских символов для русских символов.

Для возможности совместного использования латинских и русских символов применяется функция invoke_font(), которая выполняет две операции:

- закрепляет загруженный блок образов за разделом G2,
- приписывает раздел G2 к области восьмиразрядных кодов GR.

Таким образом реализуется применение восьмиразрядной кодировки КОИ-8 представления латинских и русских символов. При выдаче на терминал соответствующих кодов на экране монитора будут воспроизводиться заданные очертания символов.

4. Проблема клавиатурного ввода русских символов в восьмиразрядном представлении кода

Как уже выше отмечалось, кодировка русских символов в большинстве видов представлений - восьмиразрядная, а в клавиатурах зарубежных терминалов не предусматривается генерация таких кодов. Автором разработаны подпрограммы для обеспечения возможности ввода русских символов, при использовании стандартных семиразрядных клавиатур.

Процедуру ввода, распознавания, перекодировки и воспроизведения символов выполняют две подпрограммы-функции:

```
get_char(char_store, char_put, f_pc_koi8, f_lat_rus),
put_char(char_put,f_lat_rus).
```

В зависимости от значения флага-параметра f_pc_koi8 функция get_char выполняет перекодировку введенного символа в представление КОИ-8 или же в альтернативное PC представление и записывает результат в параметр char_store.

Параметр f_lat_rus определяет, в каком регистре (русском или латинском) вводятся символы.

В параметр char_put будет помещен код символа для выдачи на терминал в дуплексном режиме и его воспроизведения на экране. Значение этого кода зависит от способа программной знакогенерации символов.

Функция put_char выполняет непосредственно выдачу кода char_put в терминальную линию. В зависимости от значения латинско/русского регистра f_lat_rus, а также типа терминала, функция может выполнять дополнительные операции ввода/вывода. Например, для терминалов типа Televideo выдает ESC-последовательности для переключения графического режима.

При работе в латинском режиме ввода символов значения параметров функций равны:

```
char_store = char_put,   f_pc_koi8 = f_lat_rus = 0.
```

Для ввода русских символов (при выборе "русского регистра" параметром f_lat_rus) принято соглашение о соответствии клавиш с латинскими символами вводимым русским:

```
Клавиши:        a b c d e f g h i j k l m n o p q r s t u v w x y
                z [ ] | ` ~
Русские буквы:  а б ц д е ф г х и й к л м н о п я р с т у ж в ь ы
                з ш щ э ю ч
```

Большие русские буквы вводятся при нажатии этих клавиш в режиме Shift.

5. Вывод на терминал русскоязычных и смешанных текстов

Как уже отмечалось, в общем случае, выполнять непосредственно выдачу на терминал русскоязычных и, тем более, смешанных текстов из файлов штатными средствами операционных систем нельзя (даже после проведенной знакогенерации). В лучшем случае выдаваться будет какая-то абракадабра. Операцию подготовки текстовых строк для нормальной выдачи выполняет функция

```
copy_decode(string_source, string_out, f_lat_rus, f_pc_koi8).
```

Функция copy_decode, учитывая значения параметров f_pc_koi8, f_lat_rus, выполняет перепись исходного текста строки символов string_source - в одной из принятых кодировок русских символов (если они есть в данной строке) - и необходимые перекодировки либо вставки кодов переключений графических режимов (как, например, в случае терминала Televideo) в строку string_out.

Подготовленная таким образом строка может быть послана на терминал обычной операцией вывода на языке Си типа

```
printf("%s",string_out).
```

Заключение

Описанные выше методика и программное обеспечение возможности использования русских символов на зарубежных ЭВМ и терминальном оборудовании можно рассматривать как инструментарий для создания более сложных проблемных систем для обработки русскоязычных и смешанных текстовых данных. Эти разработки в полном объеме были применены автором в системе-оболочке GALAXY /1/ на ЭВМ VAX для операций просмотров файлов с различными формами представления русских символов:

- в кодировке КОИ-8,
- в альтернативной кодировке,
- в двух формах интерпретации латинских букв как русских.

Применяя систему GALAXY, можно просматривать также и MAIL-сообщения на русском языке.

Использованы эти разработки и при создании системонезависимого редактора текстов VEGA /2/.

ЛИТЕРАТУРА

1. Галактионов В.В. GALAXY - интерактивная система обслуживания файлов и процессов в ОС VMS. Общее описание возможностей и области применения. - ОИЯИ, Р11-93-24, Дубна, 1993.
2. Галактионов В. В. Системонезависимый экранный текстовый редактор VEGA. - ОИЯИ, Р11-93-22, Дубна, 1993.

Рукопись поступила в издательский отдел
26 января 1993 года.