Множественность кириллических кодировок и проблема стандартизация TEX и DVI файлов и организации работ в сетях в рамках проекта “Русский TEX”

Знаменский С. В.

http://www-sbras.nsc.ru/koi8/mathpub/znam2.html

Коды русских букв в различных операционных системах (DOS, WINDOWS, MACINTOSH, UNIX) отличаются. Хуже того, под UNIX наряду с наиболее распространенной кодировкой KOI-8 имеется международный стандарт ISO 8859-5, поддержанный новым ГОСТ. Как обеспечить пользователю возможность полноценной обработки файла, подготовленного в TEXв на другом компьютере? Как наиболее эффективно организовать структуру TEXна сервере, обслуживающем клиентов на разных платформах? Излагаются подходы к решению этих проблем, прорабатывающиеся в ходе осуществления проекта “Русский TEX” и новый набор шрифтов для TEX.

Текстовые файлы, подготовленные для обработки программой TEX, содержат разнообразные команды форматирования и активно используются физиками, математиками и учеными смежных специальностей всего мира для обмена информации и издательской деятельности. Причина популярности этого способа представления научно-технической информации в высокой компактности и читаемости файлов вне TEX, сохранении в них логической структуры документа и в полной переносимости системы TEXна любые платформы. В России распространено немало вариантов русификации, что очевидно снижает переносимость файлов.

1. Переносимость файлов

Естественно выделяются следующие уровни переносимости файлов, подготовленных в системе TEX:

Логический уровень означает трансляцию исходного файла без ошибок, вызывающих остановку TEX. Этот уровень требует лишь соответствия множества командных слов на разных машинах. Старые варианты русификаций зачастую совместимы на этом уровне. Несовместимость здесь обычно возникает из-за большого числа популярных стандартных (нерусифицированных) форматов и стилевых файлов, лишь часть из которых обычно установлены на компьютере. Если, например, в системе установлены PLAIN, AMS-TEX2.1, новый LATEX2e с AmSLATEX1.2 и оригинальный LATEX2.09, в ней возникнут проблемы с обработкой файла, подготовленного в довольно распространенном AmSLATEX1.2, которые не решатся простым добавлением стилевых файлов. Связанная с русификацией несовместимость возникает при использовании стилевых файлов, подгружающих шрифты, поскольку имена русифицированных шрифтов могут существенно отличаться. Особенно ярко это проявляется при использовании нового LATEX (LATEX2e), в котором подключение русифицированных шрифтов производится исключительно из стилевого файла. Для обеспечения совместимости на этом уровне необходимо ограничить круг используемых форматных файлов перечисленными выше, оснастить систему указанными форматными файлами и зависящими от формата путями поиска стилевых файлов (поскольку файлы с одним именем, используемые для разных форматов, не взаимозаменяемы) и фиксировать имя файла, ведающего подключением русских шрифтов и связанных с их спецификой команд в LATEX2e (предлагается russian.sty).

Совместимость на уровне TEX файла будет означает ниже идентичность разметки DVI-файла (DeVice Independent) при обработке подготовленного в TEX текстового файла. Этот более высокий уровень совместимости требует жесткой фиксации размеров букв и другой информации из TFM-файлов, способа переключения режимов переносов (с русского на английский и обратно) и тонкостей действия всех команд. Фиксация очертаний букв при этом не обязательна. Таким образом, либо должны быть идентичные стилевые файлы и шрифты (при этом лучше говорить о полной совместимости), либо стилевые файлы должны различаться лишь именами подгружаемых шрифтов, а сами шрифты должны иметь идентичную метрическую информацию. Попутно замечу, что довольно популярная идея загрузки русских и английских переносов в один язык порочна при рассмотрении на этом уровне, поскольку при такой загрузке значение `\righthyphenmin` будет несмотря на любые ухищрения оставаться одним и тем же на протяжении двуязычного абзаца (который нередок, например, в списке литературы) и слова не будут корректно переноситься.

Совместимость на уровне DVI-файлов означает, что DVI-файл, созданный на одной машине, должен обрабатываться на другой без потери и искажения содержащейся в нем информации. Поскольку информация об очертаниях конкретных букв не содержится в файле, то изменение деталей букв я не считаю искажением DVI-файла. Для совместимости на этом уровне необходимы правильные VF- или PK-файлы (если при этом есть еще и совместимость на уровне TEX файла, то можно говорить о полной совместимости), либо интеллектуальные драйверы (например, EmTEX), воспринимающие правильные TFM-файлы и алгоритм замены шрифтов.

Какой же совместимости следует добиваться, фиксируя новый стандарт? Ответ на этот вопрос далеко не очевиден. Многие, например, считают совместимость на уровне DVI-файлов излишней роскошью. На практике же немало авторов и даже издательств предлагают читателям статьи в виде DVI-файлов, поскольку это самый компактный способ их представления, не допускающий редактирования. Обходя дискуссию об этической стороне проблемы, отмечу другое немаловажное обстоятельство: многим пользователям удобно готовить DVI-файл на одном компьютере, а распечатывать – на другом. Для этого совместимость на уровне DVI-файлов уже совершенно необходима, если платить за это не придется слишком дорого. Имей каждый пользователь под рукой быстрый компьютер с гигабайтными дисками, вопрос бы вообще был риторическим, никто бы и не захотел ничего кроме полной совместимости. Но в реальной жизни все-же следует, по-видимому, искать компромисс, при котором совместимость на каждом из перечисленных уровней могла бы быть достигнута без чрезмерных дополнительных затрат ресурсов.

Рассмотрим, например, вопрос о русификации LaTeX2e. Для достижения полной совместимости нужно, чтобы файл, подгружающий русифицированные шрифты, имел имя, позволяющее однозначно идентифицировать комплект подгружаемых шрифтов. Для достижения только лишь логической совместимости (которая устроит любого разумного пользователя 286 компьютера) слишком большой роскошью было бы держать несколько вариантов русифицированных шрифтов и подгружающих их файлов. Зафиксировать один (самый первый? самый лучший? самый распространенный?) комплект шрифтов и представить остальные нестандартными может ведь и не получиться. Если не ставить в файл указание на используемый вариант русификации, не будет полной совместимости. Как же быть? Напрашивается следующий выход: договориться, что вызов файла, ведающего подгрузкой текстовых шрифтов (например, rus-xcm.sty), осуществляется командой

```
\usepackage[rus-xcm]russian
```

и при этом система, не содержащая комплекта xcm, корректно обрабатывает файл на логическом уровне совместимости, слегка посетовав в LOG-файл на невозможность воспроизвести идентичное разбиение на строки.

2. Проблема множественности кириллических кодировок

В разных операционных системах – DOS, UNIX, WINDOWS, MACOS и т.д. внутреннее представление русских букв отличается настолько,что текстовый файл, подготовленный в одной из них, не может быть прочитан в другой без перекодировки. Хуже всего дело обстоит с UNIX, где наряду со старым стандартом KOI-8 кириллической кодировки действует новый государственный и международный стандарт 8859-5. Это создает дополнительные трудности на пути достижения совместимости на уровне DVI-файла, особенно когда в локальной сети находятся клиенты на различных платформах.

Следующие подходы заслуживают внимания:

1. Фиксируется расширение CM-шрифтов русскими буквами с сохранением латинских названий. Никакого редактирования оригинальных стилевых файлов и других сложностей с русификацией. Это притягательное своей простотой решение к сожалению не свободно от серьезных недостатков. Во-первых, единственное условие, ограничивающее некоммерческое распространение этих шрифтов и их исходников состоит в том, что их нельзя модифицировать не изменяя имени, а авторское право Д. Кнута хочется уважать. А если он вдруг захочет к примеру расширить свои шрифты другим образом? Или не расширять их вовсе? И как быть нашим коллегам за рубежом? Заменять CM-шрифты на их русифицированные расширения? Эти вопросы, остающиеся без ответа, удерживают нас от следования таким путем.

Недавно автором предложен другой, уже проверенный способ избежать исправления имен шрифтов в исходных файлах, эффективный в распространенном случае, когда названия расширенных шрифтов отличаются от названий CM-оригиналов заменой букв CM на XCM или другую комбинацию. Состоит он во временном переопределении команды `\font` так, чтобы вместо латинских текстовых шрифтов грузились соответствующие русифицированные, а математические шрифты грузились стандартные.

2. Выделяется избранная кодировка шрифтов. Для корректной обработки текста как правило требуется перекодировка, которая может осуществляться либо внутри TeX (механизм TCP под DOS или OS/2 или адаптация исходников под UNIX), либо снаружи (запускающая TeX оболочка либо командный файл). Недостаток этого варианта состоит в необходимости перекодировки, которая способна внести нежелательные изменения в файл, использующий одновременно другие восьмибитные шрифты столь же явным путем. Отдельный вопрос о том, что именно следует считать некорректным: исходные файлы, использующие несколько кодовых таблиц или автоматическую перекодировку, насколько мне известно, пока не имеет общепризнанного решения среди участников проекта несмотря на очевидное соображение о том, что совсем без перекодировок нам – увы – уже не обойтись, а вот кому нужны таким образом оформленные документы на многих

языках – не очень понятно. На мой взгляд главное препятствие к использованию этого варианта состоит в том, что система должна стать достаточно интеллектуальной, чтобы у пользователя не болела голова, нужно данный файл перекодировать или же это просто приведет к потере информации. Непростой вопрос, также возникающий при таком подходе, состоит в том, какую именно кодировку фиксировать.

Предложение А.И. Роженко фиксировать в качестве базовой кодировки КОИ-8 как стандарт под UNIX принципиально не принимается РФФИ, который как государственная организация считает своим долгом поддерживать международный и государственный стандарт ISO 8859-5. На самом деле, кодировка шрифта TEX может и не совпадать ни с одной из кодировок для текстовых файлов. Более того, у нее есть достаточно веские основания чтобы не совпадать ни с одной из кодировок для текстов: текстовый шрифт содержит буквы и символы, но не может содержать акцентов без букв, а только может содержать готовые акцентированные буквы, в то время, как TEXовский шрифт совсем наоборот не обязан содержать готовых акцентированных букв.

Необходимость включения акцентированной буквы отдельным символом шрифта возникает лишь тогда, когда для языка, ее содержащего, предполагается создание и использование алгоритма автоматических переносов. В частности, если ограничиться автоматическими переносами только русских слов, можно создать TEXовский шрифт, являющийся расширением CMR10 и позволяющий включать в документ фразы на любом из 64 языков на кириллической основе кроме абхазского, в котором слишком много специфических только для этого языка букв. Есть и другие соображения, которые будут обсуждаться в связи с новым комплектом шрифтов.

3. Как можно организовать работу, не перекодируя исходный файл? Самый очевидный вариант для этого - иметь отдельные шрифты для каждой кодировки с индивидуальными названиями. Проблема лишь в том, что названий требуется очень много, что нетривиально, поскольку имя шрифта не может содержать более восьми букв и цифр и не должно совпадать с каким-либо из сотен имен уже существующих шрифтов. А.Ф. Слепухин выработал почти согласующую-

ся с международными тенденциями систему формирования имен кириллических расширений шрифтов с отведением двух символов на указание кирилличности и кодировки. Это предложение пока не получило единодушного одобрения в связи с тем, что стало трудно связывать имя стандартного шрифта с именем русифицированного, а, значит, значительно труднее русифицировать программы подключения шрифтов. Кроме того, в именах новых DC-шрифтах могут быть использованы более точные указания жирности и размера, чем в именах, предложенных А.Ф. Слепухиным.

4. Заслуживает внимания и следующая идея В.К. Малышева: входной файл обрабатывается в родной кодировке с временными шрифтами, соответствующими этой кодировке, после чего утилитой POSTEX преобразуется к стандартному набору шрифтов. Имена локальных шрифтов совпадают на разных платформах, что дает многократную экономию на именах шрифтов и избавляет от необходимости жертвовать буквами в именах и связанными с этим неудобствами. Дополнительная процедура перекодировки может быть опущена если DVI-файл не уходит из рабочей директории. Основным недостатком подхода является необходимость включения в систему средств, удерживающих оч.умелого пользователя от попыток переписать со своей машины локальные шрифты и DVI-файл, выполненный в расчете на эти шрифты, на компьютер с иной локальной кодировкой.

Сравнивая приведенные подходы, можно заключить, что ни один из них не свободен от существенных недостатков и с этим обстоятельством, похоже, придется смириться. Это означает, что выбор сделать придется невзирая на то, что какому-то кругу специалистов он все равно покажется не самым правильным.

Один из вариантов решения проблемы состоит в том, чтобы зафиксировать самое первое и наиболее распространенное в настоящее время решение, принадлежащее В.К. Малышеву. На мой взгляд, другим существенным фактором, который мог бы быть принят во внимание при выборе решения, смогла бы оказаться реальная поддержка всех языков на кириллической основе народов, проживающих на территории России, что никак не противоречит назначению РФФИ, который тем самым оказал бы неоценимую поддержку группе филологов, занимающихся языками и письменностью народов России, и заодно помог бы некоторым малым народностям обрести свою письменность.

3. Новый набор шрифтов и новые предложения по кодировкам

Наиболее широко в России распространены шрифты XCM* Самарина и Глонти, известные также под псевдонимом CMU*, русская часть называется CMCYR*, CMCZ*. и т.д. Статус этих шрифтов был не вполне ясен, поэтому в 1995 г. планировалось оснастить дистрибутив свободно распространяемыми ассоциацией пользователей кириллического TEXа шрифтами семейства LH*.*. Неожиданно для участников проекта правление ассоциации заявило о необходимости предварительных переговоров о возможности включения шрифтов в дмстрибутив с авторами шрифтов, правлением ассоциации и издательством «МИР». Зашедшие в тупик переговоры спровоцировали создание совместно с Л.Н. Знаменской альтернативного варианта шрифтов. При разработке шрифтов на наш взгляд достигнута цель сделать шрифт лучший в следующих отношениях:

1) Буквы и текст имеют более привычные для русского глаза очертания;
2) Выровнена насыщенность букв в строке;
3) Проработаны аналогов всех латинских шрифтов на основе драйверов CM;
4) Исключены ошибки, возникающие при генерации метафонтом шрифтов низкого разрешения.

Использованы базовые элементы букв семейства CM, номера контрольных точек в большинстве букв соответствуют CM и отчасти cmcyr, откуда взяты и отдельные фрагменты программ некоторых букв. Семейство LH использовалось лишь для сравнения результатов.

Вопрос о выборе базового комплекта шрифтов будет решаться на основе заключения экспертов отела математики и физики и отдела информационных систем РФФИ.

Одновременно с новыми шрифтами предложена новая внутренняя кодировка для TEX расширения кодовой таблицы (коды 128-255). Эта кодировка ориентирована на максимально полное использование возможностей EmTEX , ее достоинства проявляются только при работе под DOS. Это не лишает ее интереса, поскольку безусловно большинство пользователей TEX работают под DOS.

Речь идет об использовании механизмов TCP и `\charsubdef`. Речь о том, что на словах, содержащие акцентированные буквы, встроенный в TEX алгоритм автоматических переносов не срабатывает. Именно для распространения алгоритма автоматических переносов на содержащие акценты слова сделан примитив `\charsubdef`, реализованный в TEX 3.0 и более поздних версиях. Команды

```
\charsubdef‘\ё=xxx ‘\e
\charsubdef‘\Ё=yyy ‘\E
```

где xxx и yyy номера соответствующих акцентов, ни на что не влияют если й и Й есть в шрифте, но создают полную иллюзию наличия этих букв в шрифте в противном случае, обеспечивая автоматические переносы и подставляя в момент записи в DVI в нужное место акцентированную букву. Для TEX она останется буквой до самого момента записи в DVI файл. Таким образом, пользователь никак не сможет ощутить разницу между шрифтом, в котором присутствуют акцентированные буквы и шрифтом, в котором они созданы на свободных местах с помощью `\charsubdef`. См. подробную документацию в файлах tex4b.zip(`\emtex\doc\char_sub.doc`, `\emtex\doc\english\tex.doc`). Единственный недостаток `\charsubdef` состоит в том, что подстановка акцентов вместо букв происходит в момент записи в DVI и поэтому менять таблицу акцентов, задействованных как буквы, на ходу (в пределах документа) небезопасно, поэтому ее рекомендуется зашивать в форматный файл. Сопровождающие TEX3.0+ макросы обеспечивают независимость интерпретации команд, вызывающих акцентированные буквы от таблицы акцентов (с точность до автоматических переносов, разумеется).

Разумеется, буква “ё” не занимает столько места в таблице, чтобы ради нее одной вводить подобные усложнения. Проблема возникает лишь если мы хотим обеспечить одним комплектом шрифтов многоязычную поддержку. Если, например, поставить задачу обеспечения всех языков на кириллической основе народов, проживающих на территории России, то свободных мест для включение в кодовую таблицу акцентированных букв уже не хватит. Применение `\charsubdef` позволяет не меняя шрифт варьировать включаемые в него виртуально акцентированные буквы, то есть обеспечить полноценную поддержку дву- трех-язычных форматов на единой шрифтовой базе, что должно оказаться достаточным на ближайшую перспективу создания алгоритмов переносов для большого количества других языков. Не думаю, что это случится быстро, поскольку создать переносы это сложнейшая работа, которую могут хорошо сделать только носители языка.

Перечислю варианты форматов на базе одного комплекта шрифтов в упомянутой кодировке:

1) Автоматический перенос русского текста в том числе с ударениями и украинского с ударениями (после разработки алгоритма переносов для украинского);
2) Автоматический перенос русского текста в том числе с ударениями и татарского с ударениями (после разработки алгоритма переносов для татарского).

Если не требовать полной автоматизации переносов, в любом из перечисленных случаев обрабатываемый набором шрифтов документ может содержать текст на всех языках народов России, Украинском, Белорусском (разумеется, Болгарском), Монгольском, Тувинском, Эскимосском и языках народов Средней Азии на кириллической основе. При этом первая половина кодовой таблицы остаётся нетронутой, что позволяет включения на языках с латинской основой. Начата подготовка описания новой кодировки и ее использования с известными языками.

Руководитель проекта С. В. Знаменский