ЭЛЕКТРОННЫЕ ПУБЛИКАЦИИ И ПРОБЛЕМЫ МНОЖЕСТВЕННОСТИ КОДИРОВОК РУССКОГО ЯЗЫКА

Ю. И. Шокин, А. М. Федотов
Институт вычислительных технологий СО РАН, Новосибирск
С. И. Знаменский[1]
Институт программных средств РАН, Переславль-Залесский

1. Введение

Осознанная мировым сообществом роль информации как стратегического ресурса, стимулировала разработки новых информационных технологий как для получения и переработки больших объемов информации, так и для ее хранения и предоставления пользователям. Первое место среди новых технологий занимают сетевые информационные технологии.

Особенно велика их роль в сфере науки и образования, в связи с существующим информационным голодом, вызванным катастрофическим падением объема подписки научно-технических библиотек, существенным сокращением числа получаемых библиотеками иностранных и отечественных изданий, а также падением тиражей и числа выходящих в настоящее время русскоязычных книг и журналов. Известно, что не только развитие, но и даже поддержка научных исследований на должном уровне не мыслима без обмена информацией. Единственным выходом из создавшегося положения является использование сетевых информационных ресурсов мирового научного сообщества, представляемых сетью *Internet*, и распространение своих достижений в виде электронных публикаций.

Если в области вычислительной техники существует устойчивая тенденция роста нашего общего отставания по сравнению с развитыми западными странами, то в области телекоммуникационных и сетевых технологий, связанных с *Internet* наше отставание не так велико и оценивается сроками порядка двух-трех лет. Причем в крупных научно-образовательных центрах таких, как Москва, Санкт-Петербург, Новосибирск мы опережаем Европейские страны и развиваемся почти вровень с США. Конечно нам пока далеко до массового использования *Internet* технологий, когда каждый желающий может иметь возможность позвонить из дома и получить доступ в *Internet*. Это связано с плохим качеством наших телефонных сетей и относительной дороговизной предоставления телекоммуникационных услуг.

Однако отставание в средствах вычислительной техники, как правило не мешает пользоваться лучшими зарубежными достижениями, что и является наиболее актуальным в настоящий момент времени. Интеграция России в мировое информационное пространство со всей неизбежностью ставит проблему создания современных компьютерных сетей передачи данных, подключения их к мировым сетевым ресурсам и создания собственных информационных ресурсов, из которых одним из важнейших являются электронные публикации.

[1]Работа поддержана грантом РФФИ 95-07-19400в

Использование гипертекстовых форматов WWW (World Wide Web) очень удобно для формирования электронных публикаций. Процесс создания электронных версий текущих периодических изданий, книг, препринтов и возникновение чисто электронных изданий протекает настолько интенсивно, что через очень короткий промежуток времени популярность электронных публикаций в научном сообществе может стать очень значительной.

Уже сейчас существуют целые электронные библиотеки, в качестве примеров можно назвать электронные библиотеки Американского (`http://www.ams.org/`) и Европейского (`http://www.emis.de/`) математических обществ. Эта деятельность активно развивается и у нас, например, Издательский дом Питер (`http://www.piter-press.ru/`), Национальная служба новостей (`http://www.nns.ru/`) и др.

Информационные возможности SQL баз данных и WWW серверов весьма велики. В течение последних трех лет в мире наблюдается взрыв интереса к этим технологиям (особенно к гипертекстовым и Internet технологиям) и ежегодный многократный рост поставщиков и пользователей информационных ресурсов. Естественные свойства организации информации через WWW сервера позволяют систематизировать и удачно преподнести пользователям очень разнообразную и разноплановую информацию, включая графические образы и звук. Основным достоинством этой технологии является достаточно просто для пользователя организованная система поиска нужной информации.

2. WWW — World Wide Web

WWW - это аббревиатура от "World Wide Web" *"Всемирная паутина"* — мировая виртуальная файловая система типа hypermedia). World Wide Web является одним из самых молодых сервисов Internet: WWW родился в 1990 году в европейском исследовательском центре CERN — Европейская Лаборатория физики элементарных частиц (`http://www.cern.ch/`). В настоящее время WWW сервер Церна (`http://www.w3.org/`) является интегрирующим сервером по поддержке WWW технологий Internet. Последнюю информацию о состоянии WWW проекта можно получить на страницах World Wide Web Consortium (W3C) — (`http://www.w3.org/pub/WWW/TheProject`).

Отметим, что WWW технология является ровесником российского Internet. В 1992 году началось практическое применение этой технологии за пределами CERN, а с конца 1993 года начался колоссальный рост информационных ресурсов WWW, который привел к тому, что сегодня этот вид ресурсов Internet является самым популярным, наиболее динамично развивающимся. Отметим, что количество WWW серверов удваивается в среднем каждые три месяца, причем количество WWW серверов коммерческих организаций растет еще быстрее. Темпы роста WWW, проявляющиеся в увеличении как числа WWW серверов, так и количества доступных через WWW документов и баз данных, превышают даже весьма высокие темпы роста общего количества компьютеров и пользователей Internet.

Первые российские WWW серверы появились в начале 1994 года. Первыми Российскими WWW серверами были АО "Релком" сети EuNet/Relcom (`http://www.kiae.su/, http://www.relcom.su/`). К концу 1994 года действовало уже несколько десятков серверов, число которых к концу 1995 года о приблизилось к двумстам и продолжает неуклонно расти, например, весной 1996 года ежедневно появлялось несколько новых WWW серверов. В настоящий момент в России функционирует где-то около 700 WWW серверов, наиболее полный список российских серверов можно получить на сервере `www.mark-itt.ru`.

Первоначально WWW серверы, как и сама сеть Internet в России, создавались организациями, уже работающих в Internet или в других российских или глобальных сетях, а именно в университетах, академических и отраслевых научно-исследовательских институтах и центрах, коммерческих фирмах, являющихся сервис-провайдерами Internet. К сегодняшнему дню WWW серверы имеют государственные, коммерческие и общественные организации. Тематика WWW серверов, рассматриваемая с точки зрения информационных ресурсов непрерывно расширяется. Создаются серверы федеральных и региональных правительственных органов, общественных объединений, средств массовой информации, производственных предприятий, фирм, работающих в сфере обслуживания. Появляются электронные версии общественно-политических, научно-технических и специализированных периодических изданий. Вместе с этим возникают новые проблемы связанные с представлением специальных символов, как например математические формулы, на страницах WWW.

3. Национальная локализация

Начиная с середины 80-х кодов компьютерная подготовка документов, как для создания твердых копий и оригинал-макетов печатных изданий, так и для использования в электронном виде (передача по каналам связи или в различных информационных системах), все больше и больше вытесняет обычные способы — машинопись или типографский набор. В некоторых организациях уже давно исчезли печатные машинки и наборные автоматы — их заменили персональные компьютеры.

Постоянно совершенствуется и технология подготовки оригиналов-макетов научно-технических текстов. Такие текстовые процессоры как TEX, MS Word или PageMaker позволяют в настоящий момент подготовить для печати практически любой документ. Конечно, если речь идет о текстах содержащих математические символы пока непревзойденной является издательская система TEX, созданная Д. Кнутом. Вторым преимуществом TEX'а является то, что по структуре команд этот язык очень похож на язык HTML (HyperText Markup Language) — язык гипертекстовой разметки документов, используемый для подготовки документов для WWW. Это свойство языка TEX позволяет практически одновременно подготавливать электронную и печатную версию документов. О проблемах связанных с кодировками символов, используемых TEX'ом, и переносимостью TEX-документов мы остановимся ниже более подробно.

Вместе с компьютерным наборам возникла проблема включения в подготавливаемые тексты символов национальных алфавитов и акцентированных символов, даже в странах использующих в качестве основного английский алфавит.

Первым способом решения этой проблемы был способ передачи не самих документов, а их образов, подготовленных для печати (в формате Postscript) или просмотра (в формате Adobe PDF). Такой способ решения проблемы трудно назвать удовлетворительным, поскольку резко возрастает объем передаваемой информации (он конечно меньше чем соответствующий объем графических файлов) и требует установки на компьютер пользователя специального математического обеспечения, с одной стороны, а с другой сильно затрудняет совместную работу над документами и их дальнейшее использование.

Для большинства европейских языков, алфавиты которых созданы на основе латинского алфавита, проблема включения в электронные документы символов национальных алфавитов, более менее удовлетворительно решается с возникновением системы

UNICODE, в которой символы национальных алфавитов включаются во вторую половину кодовой таблицы (числовые коды в диапазоне от 128 до 255). Хотя для всех европейских языков одной таблицы символов не хватило.

В качестве стандарта, принятого Международной организацией стандартизации (ISO, International Standards Organisation) принято к использованию 10 различных расширений кодовой таблицы:

- Кодовая таблица ISO-8859-1, старое название Latin 1, соответствует языкам Западной и Центральной Европы: Албанский, Немецкий, Английский, Каталонский, Датский, Испанский, Финский, Французский, Фламандский, Ирландский, Исландский, Итальянский, Голландский, Норвежский, Португальский.
- Кодовая таблица ISO-8859-2, Latin 2, соответствует славянским языкам Центральной Европы и Германии: Немецкий, Хорватский, Венгерский, Польский, Румынский, Словацкий, Словенский, Чешский.
- Кодовая таблица ISO-8859-3, Latin 3, соответствует языкам: Эсперанто, Галицийский, Мальтийский, Турецкий.
- Кодовая таблица ISO-8859-4, Latin 4, соответствует языкам Восточной Европы: Эстонский, Латышский, Литовский.
- Кодовая таблица ISO-8859-5 соответствует кириллическому алфавиту, языки: Болгарский, Белорусский, Македонский, Сербский, Русский. Эта кодировка в настоящий момент является государственным стандартом, узаконенным ГОСТ'ом.
- Кодовая таблица ISO-8859-6 соответствует арабскому алфавиту.
- Кодовая таблица ISO-8859-7 соответствует греческому алфавиту
- Кодовая таблица ISO-8859-8 соответствует арабскому алфавиту.
- Кодовая таблица ISO-8859-9, Latin 5, расширение таблицы Latin 1, связанное с дополнительными буквами исландского (кельтского) языка.
- Кодовая таблица ISO-8859-10, Latin 6, другой более полный вариант кодовой таблицы для языков Восточной Европы, включая скандинавские языки.

В настоящее время ведется работа по разработке стандартов кодовых таблиц, включающих символы всех языков мира, в том числе китайского и японского.

Введение различных кодовых таблиц и локализация языковых стандартов на компьютерах (стандартные операционные системы на компьютерах таких фирм-производителей как IBM или Sun Microsystems поддерживают приведенные выше национальные языковые стандарты) полностью сняло национальную проблему использования различных текстовых процессоров для подготовки оригинал-макетов.

Однако проблема передачи электронных версий документов из страны в страну попрежнему осталась.

При передачи документов через Internet эта проблема решается путем задания расширения MIME (Multipurpose Internet Mail Extensions) и соответствующего номера кодовой таблицы (`charset=ISO 8859-1`). И если на вашем компьютере установлена операционная система поддерживающая UNICODE (или MIME), то вы уведите текст в том виде в котором его подготовил автор.

При работе с текстовыми процессорами как мы увидим это ниже на примере TEX'а дело обстоит несколько сложнее.

4. Русские кодировки

С русским кодировками дело обстоит значительно сложнее, чем с европейскими. В силу сложившихся в нашей стране традиций международный стандарт кириллического алфавита ISO 8859-5 в нашей стране очень мало применяется. Кодировка ISO 8859-5 применяется в основном на различных компьютерах с ОС UNIX, в частности на компьютерах фирмы Sun Microsystems с операционной системой Solaris. С использованием данной кодировки выполнена локализация ряда программных продуктов для UNIX-совместимых систем, например, нескольких крупных систем управления базами данных.

Кроме этой в настоящее время достаточно широко применяются четыре кодировки символов кириллицы.

KOI8-R (`charset=KOI8-R`). Кодировка, которая базируется на государственном стандарте кода обмена информацией КОИ8 и применяется в основном на компьютерах с ОС UNIX. Данная кодировка нашла отражение в документе RFC1489 (Request For Comments).

CP866 (`charset=x-CP866`). Кодовая страница (code page) Microsoft/IBM или "альтернативная" кодировка. Применяется в основном на персональных компьютерах с операционной системой MS-DOS.

CP1251 (`charset=Windows-1251`). Кодовая страница, принятая Microsoft в графических оконных системах семейства Windows и поэтому имеющая в настоящее время наиболее широкое распространение в самых разнообразных прикладных системах для IBM PC-совместимых компьютеров.

MACOS. Кодировка, используемая на компьютерах фирмы Макинтош.

В связи с тем, что документы могут храниться на серверах в различных кодировках, возникает задача корректного отображения текста программой, с помощью которой пользователь просматривает документ. Эта задача может быть решена различными способами в зависимости от того, как хранится информация на сервере, как организовано взаимодействие сервера с клиентской программой и какая программа используется для просмотра документов. Рассмотрим основные подходы, с которыми встретится пользователь, обращаясь к разным серверам.

Наиболее простой для администратора сервера способ – предоставить доступ ко всем документам в какой-либо одной кодировке. Весьма часто по этому пути идут приверженцы UNIX-систем, которые размещают файлы в кодировке КОИ8. В этом случае пользователь должен сам позаботиться об установке необходимых шрифтов и о конфигурировании своей программы просмотра. Справедливости ради следует подчеркнуть, что сделать это в большинстве случаев совсем не трудно. Графические браузеры, работающие в графических средах Microsoft Windows и XWindows System, (например, программы Netscape, Internet Explorer и Mosaic), позволяют выбрать шрифт для отображения текста документов, а соответствующие шрифты имеются для разных кодовых таблиц. Однако такой принудительный выбор кодировки имеет и некоторые неудобства. Во-первых, "скользя" по Сети и переходя с сервера на сервер, вы будете вынуждены менять настройки своего браузера. Во-вторых, при работе в среде Windows с HTML-документом в кодировке КОИ8 теряется возможность копировать фрагменты текста документа в буфер обмена и переносить их, например, в документ, созданный в текстовом процессоре MS Word с использованием "родной" кодировки CP1251. Наконец, на рабочей станции может не оказаться нужных шрифтов, а предложение скопировать их с одного из серверов Internet и установить на своем компьютере не всегда вызывает энтузиазм у пользователей, а некоторых новобранцев Internet просто приводит в состояние шока. Сейчас доля WWW серверов, на которых информация предоставляется только в одной кодировке, неуклонно уменьшается. Шрифтовые файлы для MS Windows и X Windows System имеются во многих архивах со свободным доступом, и на серверах, где поддерживается лишь одна из кодировок, как правило, указывается, где можно получить необходимые шрифты.

Диаметрально противоположный подход заключается в том, что сервер пытается сам распознать кодировку программы-клиента, после чего перекодирует результат запроса (документы) из локальной кодировки, принятой на сервере, в кодировку клиента. Для определения кодировки клиента используется возможность протокола HTTP передавать в заголовке запроса перечисление допустимых форматов документов (MIME content-type). Программа-клиент должна быть для этого должным образом сконфигурирована. В данном случае пользователь не задумывается о том, в какой кодировке хранится информация на сервере, а получает данные в кодировке, пригодной для просмотра используемым браузером.

Наиболее распространенный способ решения проблемы кодировки заключается в том, что на начальной странице WWW-сервера предлагается выбрать кодировку, в которой вы хотите получать информацию. На большинстве серверов предлагается выбор из двух кодировок – КОИ8 и CP1251 (MS Windows), кодировки CP866 (MS-DOS) и ISO-8859-5 встречаются реже. Часто для наглядности на этой странице выводят текст в нескольких кодировках, и вы сразу видите, какую из них использует ваша программа просмотра. Выполнив переход по соответствующей гиперссылке, вы в дальнейшем получаете документы в нужной кодировке. При этом на сервере может быть организовано хранение копий документов в нескольких кодировках (что требует дополнительного дискового пространства) или динамическое перекодирование документов "на лету" перед отправлением их пользователю. В целом, выбор кодировки пользователем из предлагаемого меню является неплохим решением, хотя необходимость скачивать лишнюю страницу (особенно если она насыщена графикой) лишь для того, чтобы выбрать на ней нужную кодировку, у некоторых вызывает раздражение. Иногда применяется сочетание второго и третьего подходов, когда пользователю предоставляется возможность выбора кодировки после ее автоматического определения.

Наконец, можно встретиться еще с одним способом, когда при первом обращении к серверу вы выбираете кодировку, а затем ваш выбор заносится в базу данных, в которой по IP-адресу компьютера определяется используемая на нем кодировка. При последующих обращениях к этому серверу с того же компьютера используется полученная ранее информация.

5. Текстовый процессор 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 TeX на любые платформы.

В России распространено немало вариантов русификации, что очевидно снижает переносимость файлов.

5.1. **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, однако, в системе установлены PLAIN, AMSTEX 2.1, новый LATEX 2e с AMSLATEX 1.2 и оригинальный LATEX 209, в ней возникнут проблемы с обработкой файла, подготовленного в довольно распространенном AMSLATEX 1.1, 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(LATEX 2e), в котором подключение русифицированных шрифтов производится исключительно из стилевого файла. Для обеспечения совместимости на этом уровне достаточно 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 LATEX 2e (предлагается `russian.sty`).
- Совместимость на уровне TEXфайла будет означает ниже идентичность разметки `DVI`-файла (DeVice Independent) при обработке подготовленного в TEXтекстового 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 не считается 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.
- Полная совместимость означает возможность совершенно идентичного (лишь с поправкой на качество принтера) воспроизведения документа.

Какой же совместимости следует добиваться, фиксируя 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 любом из перечисленных уровней допускалась бы и могла бы быть достигнута без чрезмерных дополнительных затрат ресурсов.

Рассмотрим, например, вопрос о русификации LaTeX 2e. 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. Попытка зафиксировать один (самый первый? самый лучший? самый распространенный?) комплект шрифтов и представить остальные нестандартными очевидно обрекла бы объявляемый стандарт на нарушение. Если же не ставить в файл указание на используемый вариант русификации, то в принципе не будет полной совместимости. Как же быть? Напрашивается следующий выход: договориться, что вызов файла, ведающего подгрузкой текстовых шрифтов (например, `rus-xcm.sty`), осуществляется явно указывающей на них командой

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

и при этом система, не содержащая комплекта `xcm`, корректно обрабатывает файл на логическом уровне совместимости, возможно слегка посетовав в LOG-файл на невозможность воспроизвести идентичное разбиение на строки. Указание на используемый пакет может быть вставлено и в закомментированный заголовок файла. На этом пути сразу возникают две непростые проблемы иного сорта:

- Запускающая TeX оболочка должна быть достаточно интеллектуальной, чтобы найти и правильно обработать информацию о варианте русификации. Проблема усугубляется большим числом операционных систем и распространенностью примитивных оболочек.
- Разработчики и администраторы TeXовских систем должны договориться между собой о формате таких указаний и их поддержке. Хорошо известно, что до сих пор подобные договоренности носили локальный характер в интересах одних групп в пику другим.

Мы вернемся к этим проблемам в конце статьи, где будет изложен конкретный путь их решения

5.2. **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 KOI8-R кириллической кодировки действует новый государственный и международный стандарт ISO 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:

- 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 одним из авторов статьи 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.
- 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 TeXоболочка 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 нам 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 в России 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, фиксировать ли кодировку и, если да, то какую именно.

  Предложение фиксировать в качестве базовой кодировки KOI8-R 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 ТеХовский 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.

- Как можно организовать работу, не перекодируя исходный файл? Самый очевидный вариант для этого - иметь отдельные шрифты для каждой кодировки с индивидуальными названиями. Проблема в том, что названий требуется очень много, что нетривиально, поскольку имя шрифта при работе под DOS 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 согласующуюся 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-шрифтов могут быть использованы более точные указания жирности и размера, чем в именах, предложенных А.Ф. Слепухиным. Еще один существенный недостаток: цифра, обозначающая кодировку, вплотную примыкает к цифрам, обозначающим размер. Другая проблема в возрастании числа TFM и MF или VF файлов, занимаемое которыми место на твердом диске с учетом кластеризации в файловой системе DOS может исчисляться десятками мегабайт, то есть составлять очень значимую часть общего объема системы.
- 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 излишне 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. Другим, на наш взгляд, 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, одним комплектом шрифтов с обеспечением этим шрифтом (под DOS, OS/2, Windows95) национальных алгоритмов переносов слов, либо тремя комплектами под остальными системами. Уместно заметить, что создание качественных алгоритмов переносов — это сложнейшая работа, которую могут хорошо сделать только носители языка.

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

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

5.3. **Наш путь решения проблемы для TeX в России.** На наш взгляд ключом к наименее болезненному для пользователей и разработчиков решению проблемы могло бы стать осуществляемое при поддержке РФФИ создание системы для обработки и создания документов в формате TEX, удовлетворяющей следующим приоритетам и требованиям:

- Компьютерная система, поддерживающая унифицированное русское расширение TeX, рекомендуемая в качестве основы для унификации, не должна сколько-нибудь существенно уступать ни одной из уже существующих установок TEX ни по качеству русифицированных шрифтов, ни по качеству переносов, ни по удобству использования, ни по уровню соответствия международным стандартам, ни по лицензионной чистоте используемых программных продуктов, ни по широте спектра предоставляемых автору документа возможностей, ни по комфорту для пользователя.
- Пользователи распространенных русификаций не должны испытывать ощутимый дискомфорт при переходе к новой системе. Файлы, подготовленные в распространенных ранее русификациях TEX, должны обрабатываться предельно корректно.
- Комфорт просмотра и распечатки документов должен быть обеспечен в максимальной степени и прежде, чем комфорт их подготовки.
- Система должна разрабатываться для наиболее распространенных в России платформ (`DOS`, `WINDOWS`) с учетом необходимости скорейшего переноса на `OS/2`, `UNIX` и другие платформы и предоставлением возможности единой файловой системы для всех платформ (`NFS`).
- Спорные проблемы, в том числе связанные с переходом от двуязычности к многоязычности должны разрешаться на основе интересов Российских пользователей с максимальным учетом тенденций и решений российских и зарубежных разработчиков и зарубежных пользователей.

Основа для такой системы (шрифты, переносы, платформонезависимая файловая структура, автоматический анализ кодировки и формата входного файла, гибкое и практичное меню с эффективной и полной системой коррекции настроек, как правило, автоматически устанавливаемых с учетом результатов анализа входного файла) уже создана. Использованный в системе алгоритм распознавания кириллической кодировки текста разработан на основе частотного анализа буквенных сочетанй в текстах на 64 языках на кириллической основе [1]. Предстоит большая работа по отбору и включению поддерживаемых пакетов, проработке многовариантной инсталляции в расчете на различные уровни совместимости, распространение на иные операционные системы и платформы и поддержка пользователей.

Подробности с конца января 1996 г. доступны по `WWW`, `URL`

`http://www.hi.botik.ru/www/tex.html`

References

[1] L.N. Znamenskaya and S.V. Znamenskii, *Russian encoding plurality problem and a new Cyrillic font set*, TUGboat, **17**(1996), no 2, 161–165.