Некоторые проблемы транслитерации и кодировки кириллицы
Some Problems of Cyrillics Transliteration and Encoding

М. Л. Халабия
Российская государственная библиотека, Москва, Россия

Mariya Khalabiya
Russian State Library, Moscow, Russia

Для упрощения процессов ведения каталога и доступа к нему пользователей зарубежные библиотеки обычно выбирают один базовый язык и один алфавит для создания библиографической записи. Таким образом, в электронном каталоге библиотеки все описания представлены максимум в двух различных системах транслитерации [1; с.3]. Под транслитерацией понимают конверсию систем письма, при которой каждый графический элемент (знак) одной системы письма представляется (заменяется) одним и тем же графическим элементом другой системы письма [2; с.2].

Поскольку в мире транслитерация в латиницу используется достаточно широко, а UNICODE до сих пор не получил достаточного распространения в информационно-библиотечных системах, перед российскими библиотеками встала проблема транслитерации кириллических символов, а также адекватное представление кириллицы в процессе международного обмена библиографическими данными. Одна из сложностей транслитерации русскоязычных библиографических и авторитетных/ нормативных записей вытекает из самой структуры русского алфавита. При орфоэпическом и фонетическом анализе символов алфавита выясняется, что русский алфавит состоит не из 33, а из 42 букв (в расчет берутся сочетания знаков), тогда как в латинице всего 26 букв [3; с.4]. Таким образом, при компьютерной кодификации не избежать диграфов и диакритических знаков.

Вторая особенность транслитерации кириллицы в латиницу заключается в существовании большого количества стандартов представления вышеназванных символов: ГОСТ 16876-71, ГОСТ 7.79-2000 (ИСО 9-95); системы транслитерации библиотеки Конгресса США и британской библиотеки и другие стандарты.

В частности, проблему выбора способа транслитерации кириллицы в латиницу РГБ решает в настоящее время при создании национального многоязычного авторитетного/нормативного файла (тезауруса) географических названий России [3]. Сейчас он содержит более 90 тыс. географических названий на русском языке. Однако в нем предполагается использовать и знаки расширенной кириллицы в названиях на государственных языках субъектов РФ. Кроме того, готовится соглашение о совместной исследовательской работе с Библиотекой Конгресса США по формированию общего файла географических названий.

UNICODE должен снять проблему представления символов различных алфавитов в одной системе и проблему международного обмена данными. Однако задачу его использования нельзя считать решенной. В электронном каталоге (ЭК) РГБ используется UNICODE (программное обеспечение АЛЕФ), но проблемы остаются.

В нашем докладе хотелось бы заострить внимание на соотношении перечня символов расширенной кириллицы, латиницы и специальных знаков, используемых в карточных каталогах Российской государственной библиотеки и закодированных для электронного каталога на старых машинах, с UNICODE.

Сравнительный анализ состава знаков расширенной кириллицы, используемой в РГБ, и Extended Cyrillic UNICODE позволил сделать следующие выводы:

1) Знаков расширенной кириллицы РГБ – 142, что составляет 57% от всех соответствующих символов UNICODE, которых насчитывается 249.

2) В каталогах РГБ практически не используются старославянские символы, которые имеются в UNICODE.

3) Кроме того, 47 диакритических символов РГБ отсутствует в стандарте UNICODE, что составляет 38% от их общего количества.

Сравнительный анализ состава знаков расширенной латиницы, используемых в РГБ, и Extended Latin UNICODE показал:

1) Знаков расширенной латиницы в каталогах РГБ – 244, символов UNICODE – 249

2) Расширенная кириллица стандарта UNICODE состоит из 3 частей:

А) C0 Controls and Basic Latin

Б) C1 Controls and Latin-1 Supplement

В) Latin Extended A

3) При этом символы расширенной кириллицы Т t; L l отсутствуют в перечне символов расширенной латиницы РГБ.

Сравнение состава специальных символов РГБ и UNICODE показало:

1) Всего специальных символов РГБ – 167. Из них 3 символа, используемых в РГБ, отсутствуют в UNICODE.

2) Главная особенность расположения символов Российской государственной библиотеке в стандарте UNICODE 4.1 – их наличие в разных таблицах системы транслитерации.

3) Необходимые таблицы символов UNICODE 4.1 для специальных символов РГБ:

1) Space separator

2) Other punctuation

3) Latin 1 Supplement

4) Letterlike symbols

5) Close punctuation

6) Open punctuation

7) Mathematical symbols

8) Other symbols

9) Currency symbols

10) General punctuation

11) Miscellaneous symbols

12) CJK symbols and punctuation

13) Mathematical Operators

14) Basic Latin

15) Spacing Modified Letter

16) Combining Diacritical Marks for symbols

17) Combining Diacritical Marks

С вышеперечисленными таблицами UNICODE можно ознакомиться на сайте: www.unicode.org/charts/normalization/index.html

Отдельно хотелось бы сказать о таблице «Combining Diacritical Marks», так как она входит в противоречие с перечнем символов расширенной кириллицы РГБ. С одной стороны, если обратиться к таблице «Extended Cyrillic», то в ней отсутствует 47 диакритических знаков, которые используются Российской государственной библиотекой. С другой стороны – наличие отдельной таблицы «Combining Diacritical Marks» дает нам эти недостающие 47 символов. На мой взгляд, кажется целесообразным ввести в таблицу «Extended Cyrillic» все буквы с этими диакритами, так как это дает возможность учесть все особенности кириллического письма.

Таким образом, стандарт UNICODE имеет средства для кодирования полного перечня символов расширенной кириллицы, латиницы и специальных знаков, используемых Российской государственной библиотекой и учитывает особенности транслитерации кириллического письма.

Литература

1. Горбатков К. К проблеме латинской транслитерации кириллицы [4; с.4]
2. ГОСТ 7.79-2000 (ИСО 9-95) Правила транслитерации кирилловского письма латинским алфавитом : Офиц. изд. / Межгос. совет по стандартизации, метрологии и сертификации. – Мн: Межгос. совет по стандартизации, метрологии и сертификации, 2002. – 19 с. [2, с.2]
3. Лавренова О.А. Национальный файл географических названий – новый проект РГБ // Библиотековедение. – 2006. – №2. – С. 46-53. [3]
4. Хохлов А.Ю. Сигла: портал доступа к библиографической информации: [Электронный ресурс]: http://www.impb.ru/~rcdl2004/cgi/get_paper_pdf.cgi?pid=34. [1, с.3].

5. www.unicode.org
6. http://www.unicode.org/charts/normalization/index.html