Reports and Working Papers

Inclusion of Nonroman Character Sets

The following document was prepared by staff of the Library of Congress as a working paper for discussions on incorporating the techniques described into the MARC communications format.

The document defines the principles for inclusion of nonroman alphabet character sets in the MARC communications format and the procedural changes needed to allow implementation of the principles. This technique was agreed upon at the MARBI Committee meeting on February 2, 1981.

Any questions on the description of the inclusion of nonroman character sets in the MARC communications format should be addressed to: Library of Congress, Processing Services, Attention: Mrs. Margaret Patterson, Washington, DC 20540.

1. INTRODUCTION

The cataloging rules followed by American libraries favor recording the title page data in the original script when possible. This helps those who consult catalogs to read the most essential information about the book. (Reading his or her name in romanized form is just as difficult for someone who knows Arabic as reading your name when it's written in Arabic.) The new cataloging rules also specify that names and titles in notes be given in their original script, AACR2 1.7A.3. Technological advances have made it possible to provide many, if not all, nonroman alphabets in machine-readable cataloging records. OCLC and RLIN are in the process of enhancing their systems so they can handle some nonroman writing systems. The Library of Congress has entered into a cooperative agreement with RLIN for the development and use of an augmented RLIN system for East Asian (i.e., Chinese, Japanese, and Korean) bibliographic data. Although the Library itself will not be creating and distributing MARC records with nonroman characters in the near term, the goal of this proposal is to define how these data can be included now so others can do so soon.

The technique known as an escape sequence announces that the codes which follow will represent letters in a specific different alphabet instead of the roman letters the codes would otherwise stand for.

2. PRINCIPLES

The following principles will govern inclusion of other alphabets in MARC records. Note that these deal only with the MARC communications format record, not the details of its processing—keying, sorting, display, etc.—by any bibliographic agency or utility. These principles are a slightly revised version of ones reviewed and approved in principle by the MARBI Character Set Committee in 1976. The earlier version was also distributed that year as working paper N77 of ISO TC46/SC4/WGI.

(1) Standard character sets should be used when available.

(2) Standard escape sequences should be used when available.

(3) Escape sequences should be used only when needed.

(4) Escape sequences are locking within a subfield but revert at any delimiter or field or record terminator code.

Example: (For demonstration purposes only, *EC* represents escape to Cyrillic and *EA* escape to ASCII)

245 10$a*EC*Russian title proper :$b*EC*Russian Subtitle. **F**

not

245 10$a*EC*Russian title proper :*EA*$b*EC*Russian subtitle. *EA***F**

and not

245 10$a*EC*Russian title proper :$bRussian subtitle.**F**

(5) Records which contain an escape sequence will also contain a special field which specifies what unusual character sets are present.

3. IMPLEMENTATION

The following will be done to realize these principles.

- The ALA character set will be redefined—see table 1.
- A new character sets present field will be defined.
- Details of application such as distribution, filing indicator values, etc., will be defined.

3.1 Discussion—ALA Character Set

A character set is a list of characters with the code used to represent each one. Using this definition, the ALA character set as given in appendixes III.B and III.C of *MARC Formats for Bibliographic Data* actually consists of eight character sets.

(1) ASCII and ALA diacritics and special characters with their eight-bit code.

(2) Superscript zero to nine, plus, minus, open and close parentheses with their eight-bit code.

(3) Subscript zero to nine, plus, minus, open and close parentheses with their eight-bit code.

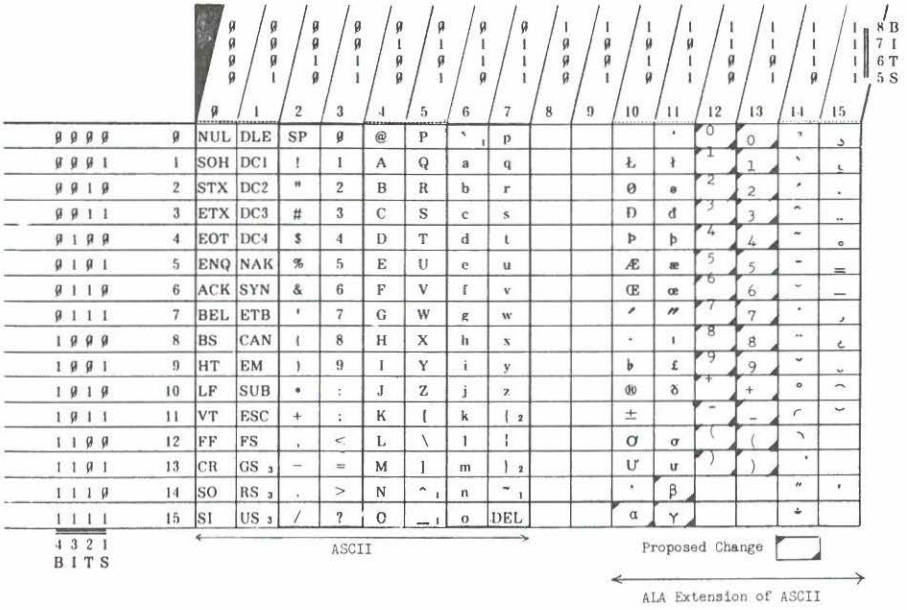
(4) Greek lowercase alpha, beta, and gamma with their eight-bit code.

(5–8) The same characters with their six-bit codes.

The six-bit character sets are used to distribute MARC records on seven-track tapes. There are very few subscribers. It is unlikely that a method can be devised for distribution of nonroman character sets records on such tapes. The present seven-track subscribers should be asked if they know of any way to do so. If they do not, the alternatives are to cease distribution of seven-track tapes entirely or limit them to those records containing only roman alphabet characters—those without a character sets present field. In the latter case, they should pay proportionately less for their subscription.

The present four eight-bit character sets and their escape sequences do not conform to present standards. The present standards did not exist when the character sets were being defined. To avoid creating and distributing records containing both standard and nonstandard character sets and stan-

Table 1. Proposed Revised ALA Character Set



dard and nonstandard escape sequences, the ALA character set should be redefined. This change will be much less traumatic than it sounds. No new characters will be added; only the codes used to represent subscript, superscript, and Greek characters will be changed. These characters were found in the title field of 8.59 out of 1.1 million records. If, as seems plausible, most or all MARC subscribers translate tapes into their own character set codes as a first step and for communication translate from their own codes into the ALA character set as the last step before distribution, only these two programs would need to be changed.

The proposed redefined ALA character set is shown in table 1. On it, columns two through seven are the American standard code for information interchange (ASCII) which is a recognized standard with a registered escape sequence. Columns ten through fifteen are the ALA extension of ASCII with special characters and the three Greek letters in columns ten and eleven, superscripts in column twelve, subscripts in thirteen, and diacritics in columns fourteen and fifteen. (It should be noted that six ASCII codes will not occur in MARC records: codes 5/14 circumflex, 5/15 underline, 6/0 grave, and 7/14 tilde are redundant with the codes for these diacritics in columns fourteen and fifteen; codes 7/11 left brace and 7/13 right brace never occur because these characters do not occur in bibliographic data. No change in this practice is proposed. It is the fact that these last two codes are used in some nonroman alphabet standard character sets that makes nonroman six-bit codes impossible.) The ALA extension of ASCII is not an official standard now; it does not have an escape sequence yet.

In addition to the ALA extension of ASCII, there is a draft international standard extended Latin alphabet character set for bibliographic use—ISO DIS 5426 (table 2). While both sets are identical in purpose, they differ in the characters they contain and the codes used to represent them. The ABACUS group has agreed that ISO 5426 be used for international distribution of MARC records among the bibliographic agencies they represent once it is an approved international standard, cf. LC *Information Bulletin*, November 16, 1979, p. 475. The Library will, however, continue to use the ALA extension for U.S. distribution. Some of the characters only on the ISO set could be added to the ALA extension without affecting existing records. An ANSI Z39 subcommittee has been established to consider this possibility. While some changes may be desirable to the ALA character repertoire, it is important that this issue not delay the separate matter of providing for inclusion of nonroman alphabets in MARC.

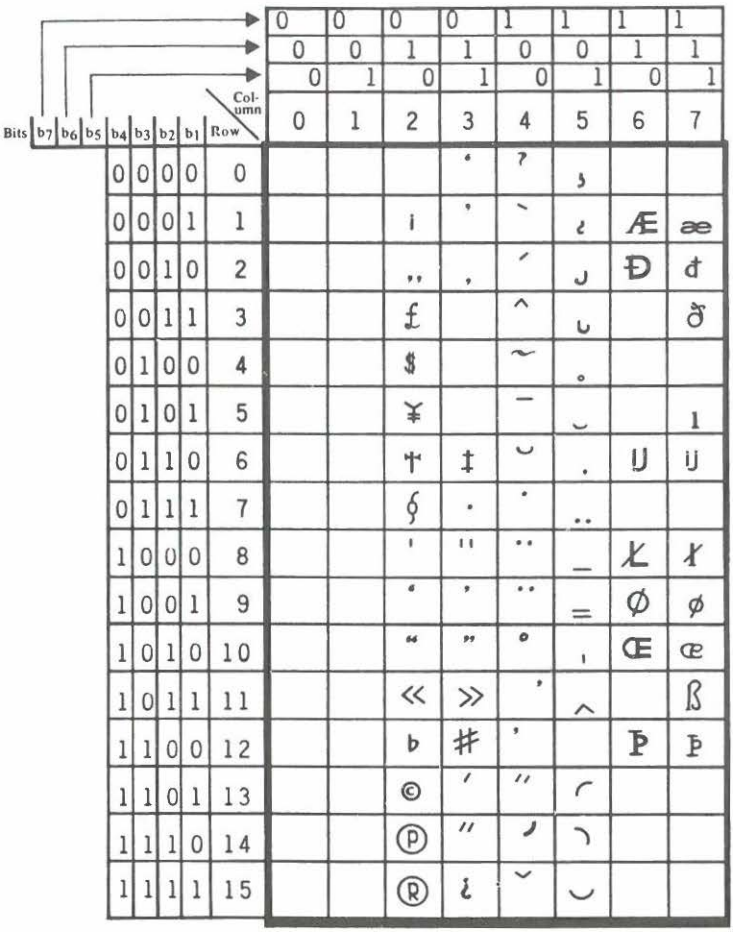
3.2 Discussion—Escape Sequence

For purposes of this discussion, escape sequences are defined as a combination of three characters. (See table 3.) The first is an escape character, hex 1/11. The second character specifies which codes are having different characters assigned to them, those in columns 2–7 or those in columns 10–15. The third character defines what characters are being assigned to these codes, e.g., Cyrillic, Greek, etc. This is a greatly simplified explanation of the escape sequence standards, ISO 2022 and ANSI X3.41. (Both are in the process of revision.) These standards provide for two types of escape sequences: public ones which reference registered character sets, and private ones for unregistered character sets. While the meaning of the latter is governed by an agreement between the sender and the receiver, they are in conformity with the standard. Until the ALA extension of ASCII has a registered escape sequence, such a "private" escape sequence could be defined for it in the character set appendix and used.

The second character of an escape sequence which changes the meaning of the codes in columns 2–7 contains either an open parenthesis, hex 2/8, or a less than sign, hex 2/12. The second character of an escape sequence which changes the meaning of the codes in columns 10–15 contains either a close parenthesis, hex 2/9, or an equal sign, hex 2/13.

The third character of escape sequences for certain registered character sets has been defined as follows:

Table 2. Extended Latin Alphabet Character Set



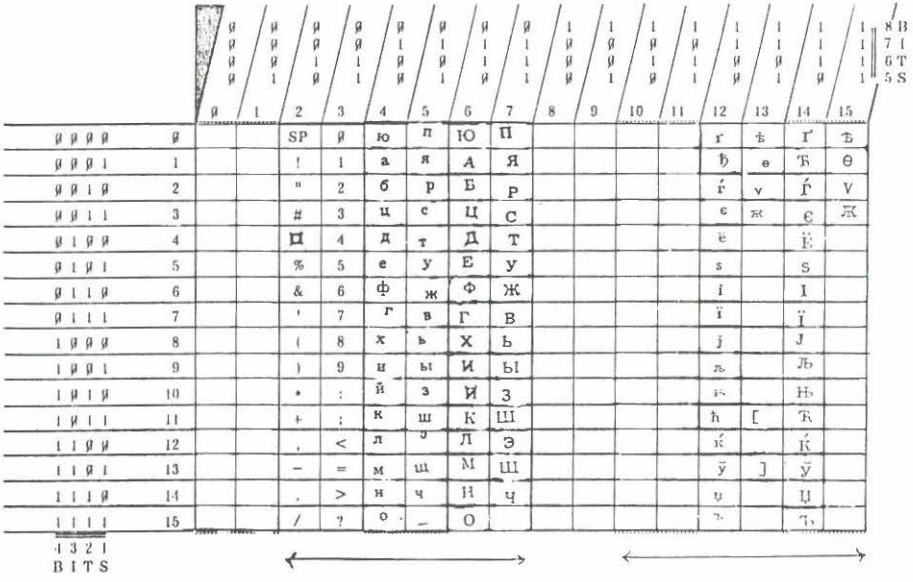
Set	*Code*
ASCII	
Russian (1967 Gost Standard) (Table 3)	registration applied for, code pending
ISO Greek	5/8, uppercase X
ISO extended Cyrillic (Table 3)	5/7, uppercase W

The sixteen codes in column three can be used to designate sixteen different "private" character sets. In MARC records, ASCII and Russian would be assigned to columns 2–7, while Greek and the extended Cyrillic (and the ALA extension of ASCII) would be assigned to columns 10–15.

Escape sequences would be given where needed in data fields. If necessary, it is permissible to embed escape sequences within a word. For example, a Latin diacritic might be needed with an extended Cyrillic letter to represent a letter in one of the non-Slavic languages of Central Asia which uses the Cyrillic alphabet.

In addition to escape sequences for nonroman alphabets described above in which one code stands for one letter, the escape standards also define escape sequence procedures for changing to multiple byte character sets. Because the ideographic writing

Table 3. Escape Sequence Character Set



systems of East Asia use thousands of different characters, it will be necessary to use two or three bytes/codes to identify a single specific character uniquely. The Japanese Industrial Standard character set, JIS 6226, uses two bytes per character, and it has been submitted to ISO to obtain a registered escape sequence. The first volume of the Chinese Character Code for Information Interchange, CCCII, has been issued; the second is expected in December. It uses three bytes per character. In all probability the LC/RLIN East Asian cooperative project will adopt either these character sets and their escape sequences or machine reversible adaptations of them. The need to expand East Asian character sets constantly to provide for infrequently used characters poses problems whose solutions cannot be predicted at this time.

3.3 Discussion—Character Sets Present Field

As specified in the sixth principle, there is need for a special field which specifies what character sets are present whenever a set other than ASCII and the ALA extension of ASCII are present in a record. The proposed field will use tag 066 and be defined as follows:

066 Character Sets Present

This field specifies what character sets are present in the other than ASCII and the ALA extension of ASCII. The field is not repeatable. Both indicators are unused and will contain blanks.

- $a This subfield will contain all but the first character of the escape sequence to the default character set in columns 2–7 whenever the default character set is not ASCII. This is not likely to occur in records created in the United States. Since there can only be one default character set, the subfield is not repeatable.
- $b This subfield will contain all but the first character of the escape sequence to the default character set in columns 10–15 whenever the default character set is not the ALA extension of ASCII. This is not likely to occur in records created in the United States. Since there can be only one default extension character set, this subfield is not repeatable.

$c This subfield will contain all but the first character (or all but the first if a longer escape sequence is used) of every escape sequence found in the record. If the same escape sequence occurs more than once, it will be given only once in this subfield. The subfield is repeatable. This subfield does not identify the default character sets.

Example:	♭♭$c)W	A record containing the ISO extended Cyrillic character set.
	♭♭$c)W$c)X	A record containing both the ISO Greek and extended Cyrillic character sets.

3.4 Discussion—Other Details

When a field has an indicator to specify the number of leading characters to be ignored in filing and the text of the field begins with an escape sequence, the length of the escape sequence will not be included in the character count.

When fields contain escape sequences to languages written from right to left, the field will still be given in its logical order. For example, the first letter of a Hebrew title would be the eighth character in a field (following the indicators, a delimiter, a subfield code, and a three-character escape sequence). The first letter would *not* appear just before the end of field character and proceed backwards to the beginning of the field.

A convention exists in descriptive cataloging fields that subfield content designation generally serves as a substitute for a space. An escape sequence can occur within a word, after a subfield code, or between two words not at a subfield boundary. For simplicity, the convention that an escape sequence does not replace a space should be adopted. One other convention is also advocated: when a space, subfield code, or punctuation mark (except open quote, parenthesis or bracket) is adjacent to an escape sequence, the escape sequence will come last.

Wayne Davison of RLIN raised the following issue. After the Library of Congress has prepared and distributed an entirely romanized cataloging record for a Russian book, a library with access to automated Cyrillic input and display capability will create a record for the same book with the title in the vernacular. (Since AACR2 says to give the title in the original script "wherever practicable," the library could be said to be obligated to do so.) In such an event the local record could have all the authoritative Library of Congress access points. To keep this record current when the Library of Congress record is revised and redistributed, it would be necessary to carry the LC control number in the local record. Most automated systems are hypersensitive to the presence of two records with the same control number. The two records can be easily distinguished: in the Library of Congress record, the modified record byte in field 008 will be set to "o" and it will not have any 066, character sets present field.

A Comparison of OCLC, RLG/RLIN, and WLN

University of Oregon Library

The following comparison of three major bibliographic utilities was prepared by the University of Oregon Library's Cataloging Objectives Committee, Subcommittee on Bibliographic Utilities. Members of the subcommittee were Elaine Kemp, acting assistant university librarian for technical services; Rod Slade, coordinator of the library's computer search service; and Thomas Stave, head documents librarian.

The subcommittee attempted to produce a comparison that was concise and jargon-free for use with the university community in evaluating the bibliographic utilities under consideration. The University Faculty Library Committee was enlisted to review this document in draft form and held three meetings with the subcommittee for that purpose. The document was also shared with library faculty and staff in order to elicit suggestions for revision.