The Structure and Content of MARC 21 Records in the Unicode Environment

Joan M. Aliprand

MARC 21 records may be encoded in individual character sets (including ASCII and ANSEL) or in Unicode (as UTF-8). This paper considers the effect of the use of Unicode without any constraints on the structure and data content of MARC 21 records. The case of Model A records where Latin is the preferred script is examined in particular detail.

Over time, the number of individual character sets that may be used in MARC 21 records has grown to twelve (see table 1).[1] Most of these character sets encode a single script, together with punctuation marks, symbols, and so on. Latin, Arabic, and Cyrillic scripts each comprise a basic and an extended character set. East Asian ideographs, Japanese katakana and hiragana, and Korean hangul are encoded in a single multi-byte character set, the East Asian Character Code (EACC).[2] Unicode equivalents have been specified for all of the characters in the individual MARC 21 character sets.[3]

Alternatively, a limited subset of Unicode characters corresponding to characters in the individual MARC 21 character sets may be used.[4] Canadian Aboriginal Syllabics may also be used in MARC 21 records encoded in Unicode.[5] The Association for Library Collections & Technical Services, Library Information & Technology Association, and Reference and User Services Association's Machine-Readable Bibliographic Information Committee (MARBI) continues to work on the technical requirements for the use of Unicode in MARC 21. A second part of the report, "Assessment of Options for Handling Full Unicode in Character Encodings in MARC 21," was discussed by MARBI at its June 2005 meeting.[6] Annex A of the report incorporates the concepts set forth in this paper, which were originally presented in June 2004 at the ALA Annual Conference.[7]

This paper examines the effect of a greatly expanded character repertoire on specific parts of MARC 21 records. What should the structure of the record be when a record is encoded in Unicode rather than in the individual character sets? In particular, how will MARC records accommodate the greatly expanded character repertoire that includes not only additional non-Roman scripts but many more Latin script characters? Are there limits on where these additional characters can be used?

Joan Aliprand is well-known as the evangelist for Unicode to the library community. She is a key player in the MARC 21 use of Unicode. Joan can be reached through Mission Professional Services at missionps@gmail.com.

Table 1. Individual MARC 21 character sets

Individual character set	Script coverage	MARC 21 category
ASCII (Basic Latin)	Latin	Default
ANSEL plus and ß (Extended Latin)	Latin	Default
Greek symbols	Greek?	Alternate: Technique 1
Subscripts		Alternate: Technique 1
Superscripts		Alternate: Technique 1
Basic Arabic	Arabic	Alternate: Technique 2
Extended Arabic	Arabic	Alternate: Technique 2
East Asian Character Code (EACC)	Ideographs, Katakana, Hiragana, Hangul	Alternate: Technique 2
Basic Cyrillic	Cyrillic	Alternate: Technique 2
Extended Cyrillic	Cyrillic	Alternate: Technique 2
Greek	Greek	Alternate: Technique 2
Hebrew	Hebrew	Alternate: Technique 2

Models for multi-script records

Historically, libraries in English-speaking areas have distinguished between Latin script and all other scripts, collectively termed "non-Roman" (or, alternatively, "non-Latin"). English is written in Latin script as are many other languages, predominantly those of Europe. Where typographical facilities for non-Roman scripts are unavailable, non-Roman text is usually rendered in Latin letter equivalents, a process called *romanization.*

To accommodate different needs worldwide, MARC 21 is flexible with respect to the structure of records containing multi-script data. Two record models, designated A and B, are specified by MARC 21.[8] A record can have data in any script in regular fields (Model B), or a record in the preferred script can be augmented with specially designated fields holding other scripts (Model A). For a particular implementation environment, Model A or Model B is chosen.

The largest use of Model A records is for MARC 21 bibliographic records with Latin as the preferred script. The use of Model A for authority records is questionable because of the complex relationships in authority data. The use of Model A for holdings data, classification data, and community information has not been explored.[9]

In figures 1 and 2, the language of cataloging is English, and the language of the monograph being cataloged is Chinese. (Text in these languages is from a record in the online catalog of the Chinese University of Hong Kong. The structural features were added by the author.)

066	##	c1
100	1#	$6880-01$aWei, Feng.
245	10	$6880-02$aJie quan dao :$bLi Xiaolong shi zhan gong fu jing cui /$cWei Feng bian zhu.
246	30	$6880-06$aLi Xiaolong shi zhan gong fu jing cui
250	##	$6880-03$aDi 2 ban.
260	##	$6880-04$aBeijing :$bBeijing ti yu da xue chu ban she,$c1992$g(1997 yin)
300	##	4, 325 p. :$bill. ;$c19 cm.
440	#0	$6880-05$aYa Zhou bo ji shu jing xuan
600	10	Lee, Bruce,$d1940-1973.
650	#0	Jeet Kune Do.
650	#0	650 0 Kung fu.
880	1#	$6100-01/$1$a 魏峰.
880	10	$6245-02/$1$a 截拳道 :$b 李小龙实战功夫精萃 /$c 魏峰编著.
880	##	$6250-03/$1$a 第 2 版.
880	##	$6260-04/$1$a 北京 :$b 北京体育大学出版社,$c1992$g(1997 印)
880	#0	$6440-05/$1$a 亚洲搏击术精选
880	30	$6246-06/$1$a 李小龙实战功夫精萃

Leader/09 = blank (i.e., individual MARC 21 character sets used)
Leader/18 = a (i.e., AACR 2 is descriptive cataloging form)

Figure 1. Significant fields of a Model A record

066	##	c1
100	1#	$aWei, Feng.
245	10	截拳道 :$b 李小龙实战功夫精萃 /$c 魏峰编著.
246	30	李小龙实战功夫精萃
250	##	第 2 版.
260	##	北京 :$b 北京体育大学出版社,$c1992$g(1997 印)
300	##	4, 325 p. :$bill. ;$c19 cm.
440	#0	亚洲搏击术精选
600	10	Lee, Bruce,$d1940-1973.
650	#0	Jeet Kune Do.
650	#0	Kung fu.

Leader/09 = blank (i.e., individual MARC 21 character sets used)
Leader/18 = a (i.e., AACR 2 is descriptive cataloging form)

Figure 2. Significant fields of a Model B record

Figure 1 shows fields from a Model A record. The preferred script for the record is Latin. The regularly tagged fields are completely romanized; Chinese written in ideographs appears in the 880 fields (alternate graphic representation). In the Model B record (figure 2) Chinese written in ideographs is in regularly tagged fields. (The subfield delimiter is shown as "$" in both records.)

In both examples, individual character sets are used (identified in the 066 field). The "$1" value in subfield c of the 066 field identifies EACC; the absence of subfields a and b from the 066 field indicates that ASCII and ANSEL are the default character sets for each record.[10]

Why two models?

In the early 1980s, Chinese, Japanese, and Korean (CJK) in their correct scripts were added to the Research Libraries Information Network (RLIN).[11] This was one of the first uses of non-Roman scripts in machine-readable bibliographic records. If the CJK data in RLIN records was to be interchanged, a decision had to be made on the inclusion of non-Roman scripts in USMARC (now MARC 21) records.

There were two options for the inclusion of CJK data in bibliographic records:

- Transcribe the CJK data directly into regularly tagged fields;
- Create a completely romanized record and append the CJK data in specially designated fields.

These options are termed Model B and Model A respectively in MARC 21.

The disadvantage of Model B was that only libraries with devices capable of displaying CJK would be able to see a complete record. In 1983, only some East Asian libraries had such devices. By contrast, the Model A option allowed any library with a system capable of displaying the characters used in ALA-LC romanization to see a complete record (although only in romanized form). When the same record was displayed on CJK-capable devices, the CJK data could be substituted for the corresponding romanization or both the original script and its romanization could be displayed.

Both Model A and Model B were specified for USMARC, and Model A was chosen for the export of records containing CJK data. Model A continues to be used for records with non-Roman scripts exported by RLG and OCLC, and in the original script cataloging files distributed by the Library of Congress.[12]

The primacy of ASCII in MARC records

"MARC 21 is an implementation of the American national standard, Information Interchange Format (ANSI Z39.2) and its international counterpart, Format for Information Exchange (ISO 2709)."[13] Each of these standards specifies a character set to encode the structural elements of the record. ANSI/NISO Z39.2 specifies use of the character set ANSI/INCITS X3.47, that is, ASCII, the acronym for "American Standard Code for Information Interchange."[14] ISO 2709 specifies ISO/IEC 646, the International Reference Version (IRV) of ASCII.[15] For brevity in the following discussion, "ASCII" should be understood to mean either ANSI/INCITS X3.47 or ISO/IEC 646 (IRV).

Because of the requirements of the ANSI/NISO Z39.2 and its international counterpart ISO 2709, the fundamental character set of any MARC record is ASCII. ASCII is also the first character set used in any MARC record because the leader, specified as the first part of an interchange record, consists of ASCII characters. Note

that an 8-bit or multi-byte character set can include characters identical to those in ASCII. For example, in the ISO/IEC 8859 family of character set standards, the first 128 characters match the characters of ISO/IEC 646.

When Unicode characters are represented in the UTF-8 encoding form, the first 128 characters of Unicode are indistinguishable from ASCII. When MARC 21 characters are encoded in UTF-8, the codes used for structural elements are consistent with ANSI/NISO Z29.2 and ISO 2709. This is why the *MARC 21 Specifications for Record Structure, Character Sets, and Exchange Media* dictate use of the UTF-8 encoding form of Unicode.[16]

MARC 21 records must be encoded either in individual character sets ("MARC-8") or in Unicode:

> If any UCS/Unicode characters are to be included in the MARC 21 record, the entire MARC record must be encoded using UCS/Unicode characters.[17]

The same character coding scheme should be used for all the records in a file: either all individual character sets or all Unicode. This restriction was recommended by the MARBI Unicode Encoding and Recognition Technical Issues Task Force in Proposal 98-18.[18]

The identification of character sets in MARC 21 records

The overall character coding scheme of a MARC 21 record is identified in character position 09 of the leader:

- # = MARC-8 (for example, individual character sets specified for MARC 21 records)
- a = UCS/Unicode (for example, the UTF-8 encoding form per the *MARC 21 Specifications*)

"UCS" is the abbreviation for "Universal Character Set," that is, ISO/IEC 10646.[19] This international standard and the Unicode Standard are kept in synch so that the character code assignments in each are identical.

When individual character sets ("MARC-8") are used, the 066 field Character Sets Present identifies all individual character sets occurring in the record (including up to two default character sets). The subfields of the 066 field are:

- a = Primary G0 character set
- b = Primary G1 character set
- c = Alternate G0 or G1 character set

The 066 field is not used in MARC 21 records encoded in UTF-8.

A MARC 21 record encoded with individual character sets (leader position 09 = [blank]) may have up to two default character sets. One of these will be ASCII, because of the requirements of ANSI/NISO Z39.2 and ISO 2709. The default character sets are identified in the nonrepeatable subfields a and b. Any other character sets occurring in the record are identified in subfield c (one occurrence for each character set).

The "G0" and "G1" designations define "default graphic character sets" in accordance with International Standard ISO/IEC 2022.[20] This standard specifies the structure of coded character sets, and how character sets are identified. The values used in the subfields of the 066 field are based on character set identifiers assigned by the International Organization for Standardization (ISO).

Additional features of character set identification

MARC 21 records that conform to Library of Congress practice have a number of additional features. The *MARC 21 Specifications* define the default graphic character sets as follows:

> ASCII graphics are the default G0 set and ANSEL graphics are the default G1 set for MARC 21 records. . . . These are the default working sets for data transcribed in the fields and subfields unless other default sets are specified in the record field 066 (Character Sets Present).[21]

ANSEL is the acronym for "American National Standard Extended Latin," that is, a shortened form of the title of the American National Standard ANSI/NISO Z39.47. The alignment of the MARC 21 Basic and Extended Latin sets with ANSEL was published in 1994.[22] The MARC 21 Extended Latin character set is now a superset of ANSEL; the euro sign and the Eszett were added in 2002 to achieve harmonization with UKMARC.[23]

As shown in the right-most column of table 1, the *MARC 21 Specifications* categorize the individual character sets as default or as alternate. Alternate graphic character sets are identified according to the technique by which they are invoked for use:

- Technique 1, used for the Greek symbols, subscript, and superscript characters, is specific to MARC 21;
- Technique 2, used for all other alternate character sets, is an implementation of ISO/IEC 2022.

In MARC 21 records consistent with Library of Congress practice, the 066 field is included only when one or more of the alternate character sets of the Technique 2 category occur in the record. The default character sets, ASCII and ANSEL, are taken as given; subfields a and b are not included in the 066 field.

Control codes in MARC 21 records

In 8-bit single-byte character sets that conform to ISO/IEC 2022, two areas are designated for control codes:

hex values 00 to 1F (C0 control codes) and hex values 08 to 9F (C1 controls). While all of these code points can be mapped to equivalent Unicode code points, the Unicode Standard specifies semantics for only ten of the code points.[24]

The *MARC 21 Specifications* define two sets of control code positions that are consistent for all records. The two sets are designated C0 and C1 in accordance with ISO/IEC 2022 conventions. Only eight of the control code positions are actually used. The C0 set has the control characters Escape, File terminator, Record terminator, and Subfield delimiter. These control codes have the same functionality as the control codes defined at the corresponding code points in ISO/IEC 6429.[25]

The C1 area has the control characters Non-sort begin, Non-sort end, Joiner, and Non-joiner, which are all specific to MARC 21. The MARC 21 characters Non-sort begin and Non-sort end are analogous to the control characters Non-sorting character(s), beginning (NSB) and Non-sorting character(s), end (NSE) at the same code positions in ISO 6630.[26]

Table 2. MARC 21 control codes and their Unicode equivalents

MARC 21 control code	Unicode equivalent	Unicode semantic
1B Escape	U+001B	*None*
1D File terminator	U+001D	Information separator three (GS)
1E Record terminator	U+001E	Information separator two (RS)
1F Subfield delimiter	U+001F	Information separator one (US)
08 Non-sort begin	U+0098	*None*
09 Non-sort end	U+009C	*None*
0D Joiner	U+200D	ZERO WIDTH JOINER
0E Non-joiner	U+200C	ZERO WIDTH NON-JOINER

MARC 21 control codes and Unicode

Table 2 shows the MARC 21 control codes and their Unicode equivalents (as the Unicode scalar values that are used in the published Unicode code charts). The right-most column shows which Unicode code point is assigned a semantic value in the Unicode Standard. In the code tables published on the MARC 21 Web site, the mappings are shown as both the Unicode scalar value and the byte sequence representing the Unicode character when the encoding form is UTF-8.

The MARC 21 control characters 1D File terminator, 1E Record terminator, and 1F Subfield delimiter are functionally equivalent to the ISO/IEC 6429 control characters with the same code points: Information separator three (also known as Group separator or GS), Information separator two (Record separator or RS), and Information separator one (Unit separator or US). These MARC 21 control characters are mapped to Unicode characters whose semantics are consistent with the corresponding characters of ISO/IEC 6429.

The MARC 21 control characters 8D Joiner and 8E Non-joiner are functionally comparable to the Unicode characters zero width joiner and zero width non-joiner. Both pairs of characters have a common origin: Xerox Corporation's implementation of Arabic script.[27]

The remaining MARC 21 control characters (1B Escape, 88 Non-sort begin, and 89 Non-sort end) map to Unicode code points that have no specified Unicode semantic. The MARC 21 control character Escape is irrelevant when MARC 21 records are converted to Unicode. The Unicode Standard states, "No escape sequence or control code is required to specify any character in any language."[28]

The interpretation of the Unicode code points equivalent to the MARC 21 control characters 88 Non-sort begin, and 89 Non-sort end is outside the scope of the Unicode Standard. If general-purpose software from original equipment manufacturers (OEMs) interprets these code points at all, it is likely to do so in accordance with the functionality specified in ISO/IEC 6429 for the characters 98 Start of string and 9C String terminator.

The *MARC 21 Specifications* must define a higher-level protocol that dictates how the Unicode code points equivalent to the MARC 21 control characters 88 Non-sort begin and 89 Non-sort end are to be interpreted, and require use of the protocol. Vendors of library systems will have to create the software that supports this higher-level protocol.

For consistency with UNIMARC, the functionality of the higher-level protocol should be fully consistent with what ISO 6630 specifies for the control characters NSB and NSE. (This will also allow vendors to use the same application software for MARC 21 and UNIMARC data.) Guidelines on use of the higher-level protocol in MARC 21 can be promulgated if any temporary limitations on the use of the two control characters are needed.

Characters allowed in components of the MARC 21 record

The fundamental character set in MARC 21 records is ASCII, because it is "used for the structural elements

of the record, and most coded data are also specified within the ASCII range of characters."[29] For example, "Only ASCII graphic characters are allowed in the Leader."[30]

Table 3 shows the components of the MARC 21 record and what is allowed in each component when the record is encoded with individual character sets. A component name ending with "field data content" means the data content of that type of field, excluding the tag, indicators, all subfield pairs, and the field terminator.

In table 3, Model A and Model B are identical, except for the content of regular and 880 fields. The 880 field contains "The fully content-designated representation, in a different script, of another field in the same record."[31] The 880 field cannot contain only preferred scripts (which can be used in regular fields); it must contain at least one additional script.

Table 4 shows the same record components, and what is allowed in each component when the record is encoded according to the current MARC 21 specifications for use with UTF-8.

When individual character sets are used in a Model A record, use of only preferred scripts in the regular fields can be enforced by allowing only certain character sets to be used for data entry in the regular fields. Both OCLC and RLIN21 do this.[32] How will this restriction be enforced when the Model A record is encoded using Unicode (as UTF-8)?

Table 3. Individual character sets and the components of the MARC 21 record

Component	Model A record	Model B record
Leader	ASCII	ASCII
Directory	ASCII	ASCII
Tags	ASCII	ASCII
Indicators	ASCII	ASCII
Subfield delimiter	Control character 1F (Unit separator)	Control character 1F (Unit separator)
Subfield code	ASCII	ASCII
Regular field data content	Character set(s) for preferred script(s)	Character set(s) for one or more scripts
880 field data content	Character set(s) for one or more scripts. At least one alternate (non-preferred) script must be present.	*not applicable*
Field terminator	Control character 1E (Record separator)	Control character 1E (Record separator)
Record terminator	Control character 1D (Group separator)	Control character 1D (Group separator)

Table 4. Characters in UTF-8 and the components of the MARC 21 record

Component	Model A record	Model B record
Leader	ASCII range only	ASCII range only
Directory	ASCII range only	ASCII range only
Tags	ASCII range only	ASCII range only
Indicators	ASCII range only	ASCII range only
Subfield delimiter	U+001F i.e., hex 1F in UTF-8	U+001F i.e., hex 1F in UTF-8
Subfield code	ASCII range only	ASCII range only
Regular field data content	Preferred script(s)	One or more scripts
880 field data content	One or more scripts. At least one alternate (nonpreferred) script must be present.	*not applicable*
Field terminator	U+001E i.e., hex 1E in UTF-8	U+001E i.e., hex 1E in UTF-8
Record terminator	U+001D i.e., hex 1D in UTF-8	U+001D i.e., hex 1D in UTF-8

A closer look at the Model A record

In the introduction to the section on character sets, the *MARC 21 Specifications* describe MARC 21 records encoded using the individual character sets:

> All content designation in MARC 21 records is encoded using the repertoire found in Code for Information Interchange (ASCII) (ANSI X3.4) or its international counterpart, ISO 646 (IRV). Other character repertoires such as the Extended Latin Alphabet Coded Character Set for Bibliographic Use (ANSEL) (ANSI Z39.47) and MARC 21 character codes for 14 superscript characters, 14 subscript characters, and 3 Greek symbols are commonly used in records with Latin script data content. Additional characters [sic] repertoires for the Arabic, Chinese, Cyrillic, Greek, Hebrew, Japanese, and Korean scripts have been designated for use in MARC 21 records.[33]

The MARC 21 Extended Latin character set is now a superset of ANSEL after the addition of the euro sign and Eszett in 2002.

Table 5 shows the components and individual character sets of a Model A record when the record conforms to the practice described in the above quotation. When Latin is the preferred script (as is the case in records that follow Library of Congress cataloging practice), data in non-Roman scripts is confined to 880 fields. (The three Greek symbols are regarded as symbols, not as Greek letters.)

Note that the character sets "commonly used for Latin script data content" match what the Program for Cooperative Cataloging (PCC) calls the *Latin base:*

> Program records are encoded in a basic complement of character sets referred to in these guidelines as the "Latin base" (ASCII, ANSEL, MARC 21 Greek, MARC 21 subscript, MARC 21 superscript).[34]

The inclusion of non-Roman data in a PCC record is optional.

The term "ANSEL" in the above definition is presumably intended to mean "MARC 21 Extended Latin characters," rather than only the characters encoded in ANSI/NISO Z39.47, that is, to include the euro sign and Eszett as well. "MARC 21 Greek" means the three Greek symbols, not the MARC 21 Greek character set with the complete Greek alphabet.

What happens when a Model A record with Latin base characters in the regular fields and non-Roman characters is converted to Unicode? Table 6 shows MARC 21 individual character sets and characters and the Unicode character blocks that contain the mapped equivalents.

The Escape character (in the C0 controls row) is included only because it is part of the "escape sequences" that identify individual MARC 21 character sets. The corresponding Unicode code point should not be used in MARC 21 records. One of the benefits of Unicode is that multi-script data can be encoded without the need for escape sequences.

The Basic Arabic, Basic Cyrillic, and Hebrew character sets include characters that also occur in ASCII. The mapping of these characters to the C0 Controls and Basic Latin character block is omitted from table 5 because the characters in question are essentially ASCII "clones."

How can the data content of such a Model A record be replicated in the context of Unicode? In particular, how can the characters that are "commonly used for Latin script data content" be identified?

Greek letters as "Latin script data content"

With the individual MARC 21 character sets, the Greek symbols (used in scientific terms, for example) can be distinguished from letters of the Greek alphabet in the MARC 21 Greek character set because each character set is uniquely identified. When the Greek symbols are converted to Unicode, there is no way to distinguish them from the regular Greek letters to which the Greek character set is mapped. Table 6 shows the Greek and Coptic character block under "Latin script data content" and under "non-Roman character sets" as well.

The way to resolve this conundrum is to follow the Library of Congress Rule Interpretation (LCRI) on how to transcribe Greek letters that occur in sources of information. Because the three symbols normally occur as isolated Greek letters, they should be transcribed as the English name of the letter enclosed in square brackets. This solution is now recommended in the MARC 21 Code Table Greek Symbols.[35]

Table 5. Occurrence of individual MARC 21 character sets in components of the Model A record

Component	Character sets
Leader	ASCII
Directory	ASCII
Tags	ASCII
Indicators	ASCII
Subfield delimiter	Control character 1F (Unit separator)
Subfield code	ASCII
Regular field data content	At least one of ASCII, ANSEL, Greek symbols, subscripts, or superscripts. Any other of these character sets may also be used.
880 field data content	At least one of the MARC 21 non-Roman character sets. Any other MARC 21 character set may also be used.
Field terminator	Control character 1E (Record separator)
Record terminator	Control character 1D (Group separator)

Extending the scope of "Latin script data content"

Many more letters from Latin script alphabets are encoded in Unicode than are available in the individual MARC 21 character sets; for example, the ij ligature encoded in ISO 5426 or the African letters of ISO 6438.[36]

But the Latin script data content does not only consist of Latin letters. The Unicode character blocks Spacing

Table 6. Unicode character blocks with characters mapped from individual MARC 21 character sets or subsets

Individual MARC 21 character sets or subsets	Unicode character blocks containing mapped equivalents
Control characters	
C0 controls: Escape; File terminator; Record terminator; Subfield delimiter	C0 Controls and Basic Latin
C1 controls: Non-sort begin; Non-sort end; Joiner; Non-joiner	C1 Controls and Latin-1 Supplement; General Punctuation
Space characters	
Space	C0 Controls and Basic Latin
CJK Space	CJK Symbols and Punctuation
Latin script data content	
ASCII (Basic Latin)	C0 Controls and Basic Latin
ANSEL plus € and ß (Extended Latin)	C1 Controls and Latin-1 Supplement; Latin Extended-A; Latin Extended-B; Spacing Modifier Letters; Combining Diacritical Marks; Currency Symbols; Letterlike Symbols; Miscellaneous Symbols
Greek symbols	Greek and Coptic
Subscripts	Superscripts and Subscripts
Superscripts	C1 Controls and Latin-1 Supplement; Superscripts and Subscripts
Non-Roman character sets	
Basic Arabic	Arabic
Extended Arabic	Arabic
East Asian Character Code (EACC)	CJK Symbols and Punctuation; Hiragana, Katakana, Hangul Compatibility Jamo; CJK Unified Ideographs; CJK Unified Ideographs Extension A; CJK Unified Ideographs Extension B; Hangul Syllables; Halfwidth and Fullwidth Forms; Private Use code points
Basic Cyrillic	Cyrillic
Extended Cyrillic	Cyrillic
Greek	Greek and Coptic
Hebrew	Hebrew; Alphabetic Presentation Forms

Modifier Letters, Combining Diacritical Marks, Currency Symbols, Letterlike Symbols, and Miscellaneous Symbols all include equivalents for ANSEL characters. When any Unicode character can be used in a MARC 21 record, should every character from these blocks be regarded as Latin script data content? Does the addition of any new characters to these blocks automatically extend the scope of Latin script data content?

But why these blocks and not other comparable ones? Why Letterlike Symbols but not Math Alphanumeric Symbols that contains letterlike symbols used in mathematical notation? If Math Alphanumeric Symbols, why not Mathematical Operators and other blocks containing characters related to mathematics? If Miscellaneous Symbols, why not Dingbats or Miscellaneous Technical? It is extremely difficult to draw the border for Latin script data content by examining character blocks, and examining individual characters would be too time-consuming.

Is there any other way to specify Latin script data content for MARC 21 records encoded in Unicode? A relatively clean way is to use the Script Name property defined in the Unicode Standard. Every Unicode character is assigned one—and only one—Script Name property, described in Unicode Standard Annex #24.[37] The values for the Script Name property are the names of specific scripts (for example, *Armenian*, *Canadian_Aboriginal*, or *Hangul*) plus *Common* and *Inherited*.

Characters with the Script Name property *Common* may be used with multiple scripts. Unassigned code points also have the *Common* property. Examples of characters having the Script Name property *Common* are:

- punctuation marks used with more than one script;
- digits used with more than one script;
- currency symbols; and
- general purpose symbols such as percent sign or music natural sign.

A character with the Script Name property *Inherited* inherits its script from its "base character," that is, from the character with the Script Name property other than *Inherited* that precedes it. Examples of characters with the *Inherited* property are: combining diacritical marks, Arabic harakat, and the zero width joiner and zero width non-joiner.

Table 7 shows that, if the three Greek symbols are converted according to LCRI, all of the characters currently allowed in a regular field have one of three Script Name properties: *Latin, Common,* or *Inherited.* In the case of *Inherited,* the preceding base character would have to have the Script Name property *Latin* or *Common.* An edit to allow all Unicode equivalents for the characters of ASCII, the ANSEL superset, superscripts, and subscripts into the regular fields while excluding characters from other scripts (for example, with a Script Name property such as *Cyrillic* or *Han*) would require the incoming character to meet one of these conditions:

- character has Script Name property = *Latin* | *Common*
- character has Script Name property = *Inherited*, and its base character has Script Name property = *Latin* | *Common*

Table 7. Individual MARC 21 character sets or subsets and the Unicode Script Name properties of the equivalent Unicode characters

Individual MARC 21 character sets or subsets	Script Name properties of Unicode mapped characters
Control characters	
C0 controls: Escape; File terminator; Record terminator; Subfield delimiter	Common
C1 controls: Non-sort begin; Non-sort end; Joiner; Non-joiner	Common, Inherited
Space characters	
Space	Common
CJK Space	Common
Latin script data content	
ASCII (Basic Latin)	Common, Latin
ANSEL plus € and ß (Extended Latin)	Common, Inherited, Latin
Greek symbols	Common, Latin (when LCRI is followed)
Subscripts	Common
Superscripts	Common
Non-Roman character sets	
Basic Arabic	Common, Inherited, Arabic
Extended Arabic	Inherited, Arabic
East Asian Character Code (EACC)	Common, Hiragana, Katakana, Katakana_ Or_Hiragana, Han, Hangul
Basic Cyrillic	Common (all ASCII clones), Cyrillic
Extended Cyrillic	Common (all ASCII clones), Cyrillic
Greek	Common, Inherited, Greek
Hebrew	Common, Hebrew

Because every Unicode character has a specific Script Name property, determining Latin script data content from the Script Name property has additional advantages:

- The scope of Latin script data content in MARC 21 records is defined for all Unicode characters without laborious examination, time-consuming discussion, and possible errors;
- The Unicode Consortium maintains a data file for the Script Name property;
- Edits to limit fields to Latin script data content will apply to all Unicode characters;
- New characters added to Unicode will be automatically covered by such edits (only the controlling data file containing the Script Name property for characters will need to be updated in software).

The question of Braille

The Unicode Standard encodes the complete set of the eight-dot patterns of Braille. The Script Name property of each of these 256 characters is *Braille.* The assignment of meanings to Braille patterns is outside the scope of the Unicode Standard.

The role of Braille as a script in a MARC 21 record has to be decided. English and other languages written in Latin script can be represented in Braille, but so can languages written in other scripts. Should Braille be included in Latin script data content, or should it be regarded as alternate graphic representation? There does not appear to be a strong reason to add Braille patterns to the scope of Latin script data content.

Why Latin script data content should be edited

The *MARC 21 Specifications* describe the data content of MARC 21 records encoded using the individual MARC 21 character sets (quoted at the beginning of the section "A

closer look at the Model A record"). Such a record can be used by any library system that conforms to MARC 21 as implemented by the Library of Congress. In such records, Latin is the preferred script, and the regular fields are limited to the individual character sets described as "commonly used in records with Latin script data content."

For records that conform to Library of Congress practice, characters outside the definition of Latin script data content should not be permitted in the regular fields. Data loading into systems that restrict regular fields to Latin script data content will be more affected when incoming records include characters outside the Latin script data content defined for the regular fields.

Latin script content and authority records

In the Model A record where Latin is the preferred script, access points tagged as regular fields are limited to Latin script data content (equivalent to the PCC term *Latin base*). When an access point limited to Latin script data content is under authority control, the established form in the 1XX field of the authority record must also be limited to Latin script data content. In addition, 5XX (See Also From tracings) fields must be limited to Latin script data content, because the 5XX fields identify different authorized forms of headings related to the authorized form of heading in the 1XX field.

Effect on conversion to MARC-8

Not all systems will be able to handle Unicode initially. Therefore, it must be possible to convert a Unicode record to the individual MARC 21 character sets ("MARC-8"). At a minimum, it should be possible to convert the data in the regular fields to ASCII and ANSEL plus the subscripts and superscripts. The Greek symbols will not be converted back, remaining as the letter names enclosed in brackets.

Converting from Unicode to individual MARC 21 character sets will be simplified if the regular fields in Unicode-encoded records are restricted to Latin script data content. With the restriction in place, it will only be necessary to provide substitutions for punctuation, digits, symbols, and Latin script letters outside the range of ASCII, ANSEL, and the superscript and subscript character sets. With no restriction on what characters may appear in the regular fields, substitutes in regular fields would be more common.

Conclusion

For a Model A record encoded in Unicode with Latin as the preferred script, the definition of the Latin script data content of regular fields should be controlled by the Script Names property of Unicode. The Script Names property values *Latin*, *Common*, and *Inherited* can be used to identify the characters that constitute Latin script data content. To create consistent records for MARC 21 interchange, the data content in Model A records with Latin as the preferred script should be controlled so that only characters consistent with the definition of Latin script data content are permitted in regular fields.

[*Author's note:* the Unicode Consortium is a registered trademark, and Unicode is the trademark of Unicode, Inc. All other trademarks are the property of their respective owners.]

References

1. Library of Congress Network Development and MARC Standards Office, "Character Sets: Part 1—MARC-8 Environment. Accessing Alternate Graphic Character Sets," *MARC 21 Specifications for Record Structure, Character Sets, and Exchange Media*, Jan. 2000. Accessed Mar. 28, 2005, www.loc.gov/marc/specifications/specch armarc8.html#alt.

2. National Information Standards Organization, *East Asian Character Code for Bibliographic Use* (Bethesda, Md.: NISO Pr., 1989). ANSI/NISO Z39.64 - 1989(R2002).

3. Library of Congress Network Development and MARC Standards Office, "Character Sets: Part 3—Code Tables," *MARC 21 Specifications for Record Structure, Character Sets, and Exchange Media*, Sept. 2004. Accessed Mar. 28, 2005, www.loc.gov/marc/specifications/ specchartables.html.

4. Library of Congress Network Development and MARC Standards Office, "Character Sets: Part 2—UCS/Unicode Environment," *MARC 21 Specifications for Record Structure, Character Sets, and Exchange Media*, June 2003. Accessed Mar. 28, 2005, www.loc .gov/marc/specifications/speccharucs.html.

5. Canadian Committee on MARC, "Proposal 2002-11: Repertoire Expansion in the Universal Character Set for Canadian Aboriginal Syllabics," May 9, 2002. Accessed Mar. 28, 2005, www .loc.gov/marc/marbi/2002/2002-11.html.

6. Library of Congress Network Development and MARC Standards Office, "Assessment of Options for Handling Full Unicode in Character Encodings in MARC 21. Part 2: Issues," June 2005. Accessed Sept. 21, 2005, http://www.loc.gov/marc/ marbi/2005/2005-report01.pdf.

7. Joan M. Aliprand, "True Scripts in Library Catalogs—The Way Forward." Presented at the program "Library Catalogs and Non-Roman Scripts," 2004 ALA Annual Conference, Orlando, Fla. Accessed Sept. 22, 2005, http://www.ala.org/ala/alcts/alctsconted/ presentations/Aliprand2.pdf.

8. Library of Congress Network Development and MARC Standards Office, "MARC 21 Concise Format for Bibliographic Data. (2004 Concise Edition)—Multiscript Records," Feb. 11, 2005. Accessed Mar. 28, 2005, www.loc.gov/marc/bibliographic/ ecbdmulti.html.

9. Joan M. Aliprand, "Scripts, Languages, and Authority Control," *Library Resources & Technical Services* 49, no. 4 (2005): 243–49.

10. American National Standards Institute, *Information Systems—Coded Character Sets—7-Bit American National Standard Code for Information Interchange (7-bit ASCII)*. (New York: American National

Standards Institute, 1986). ANSI/INCITS 4-1986(R2002), formerly ANSI X3.4-1986(R1997); National Information Standards Organization, *Extended Latin Alphabet Coded Character Set for Bibliographic Use* (Bethesda, Md.: NISO Pr., 1993). ANSI/NISO Z39.47-1993(R2002).

11. John W. Haeger, "The CJK Enhancements to the RLIN System: a Review of Basic Issues," in *Automated Systems for Access to Multilingual and Multiscript Library Materials: Problems and Solutions: Papers from the Preconference Held at Nihon Daigaku Kaikan, Tokyo, Japan, Aug. 21–22, 1986,* ed. for the Section on Library Services to Multicultural Populations and the Section on Information Technology by Christine Bossmeyer and Stephen W. Massil (Munich: Saur, 1987), 156–62.

12. Library of Congress Cataloging Distribution Service, "LC Original Script Cataloging Files," Jan. 12, 2005. Accessed Mar. 28, 2005, www.loc.gov/cds/mds.html#nrcf.

13. Library of Congress Network Development and MARC Standards Office, "Record Structure," *MARC 21 Specifications for Record Structure, Character Sets, and Exchange Media*, Jan. 2000. Accessed Mar. 28, 2005, www.loc.gov/marc/specifications/specrecstruc.html.

14. National Information Standards Organization, *Information Interchange Format* (Bethesda, Md.: NISO Pr., 1994). ANSI/NISO Z39.2 – 1994; American National Standards Institute, *American National Standard Code Extension Techniques for Use with the 7-Bit Coded Character Set of American National Standard Code for Information Interchange* (New York: American National Standards Institute, 1974). ANSI X3.41-1974.

15. International Organization for Standardization, *Information and Documentation—Format for Information Exchange* (Geneva: International Organization for Standardization, 1996). ISO/IEC 2709:1996; International Organization for Standardization, *Information Technology—ISO 7-Bit Coded Character Set for Information Interchange* (Geneva: International Organization for Standardization, 1991). ISO/IEC 646:1991.

16. Library of Congress Network Development and MARC Standards Office, "Character Sets: Part 2—UCS/Unicode Environment, Specification, Encoding Rules," *MARC 21 Specifications for Record Structure, Character Sets, and Exchange Media*, June 2003. Accessed Mar. 28, 2005, www.loc.gov/marc/specifications/speccharucs.html.

17. Library of Congress Network Development and MARC Standards Office, "Character Sets: Part 2—UCS/Unicode Environment. Specification. UCS/Unicode Markers and the MARC 21 Record Leader," *MARC 21 Specifications for Record Structure, Character Sets, and Exchange Media*, June 2003. Accessed Mar. 28, 2005, www.loc.gov/marc/specifications/speccharucs.html.

18. MARBI Unicode Encoding and Recognition Technical Issues Task Force, "Proposal 98-18: Unicode Identification and Encoding in USMARC Records," May 27, 1998. Accessed Mar. 28, 2005, www.loc.gov/marc/marbi/1998/98-18.html.

19. International Organization for Standardization, *Information Technology—Universal Multiple-Octet Coded Character Set (UCS)* (Geneva, International Organization for Standardization, 2003). ISO/IEC 10646:2003.

20. International Organization for Standardization, *Information Technology—Character Code Structure and Extension Techniques* (Geneva: International Organization for Standardization, 1994). ISO/IEC 2022:1994, and *Corrigendum 1* (Geneva: International Organization for Standardization, 1999).

21. Library of Congress Network Development and MARC Standards Office, "Character Sets: Part 1—MARC-8 Environment. Graphic Character Sets," *MARC 21 Specifications for Record Structure, Character Sets, and Exchange Media.* Jan. 2000. Accessed Mar. 28, 2005, www.loc.gov/marc/specifications/speccharmarc8.html.

22. Library of Congress Network Development and MARC Standards Office, *USMARC Specifications for Record Structure, Character Sets, and Exchange Media,* 1994 ed. (Washington, D.C: Cataloging Distribution Service, Library of Congress, 1994).

23. British Library et al., "Proposal 2002-14: Changes for UKMARC Format Alignment," May 21, 2002. Accessed Mar. 28, 2005, www.loc.gov/marc/marbi/2002/2002-14.html.

24. *The Unicode Standard, Version 4.0* (Boston: Addison-Wesley, 2003), 386.

25. International Organization for Standardization, *Information Technology—Control Functions for Coded Character Sets* (Geneva: International Organization for Standardization, 1992). ISO 6429:1992.

26. International Organization for Standardization, *Documentation—Bibliographic Control Characters* (Geneva: International Organization for Standardization, 1986). ISO 6630:1986.

27. Joseph D. Becker, "Arabic Word Processing," *Communications of the ACM* 30, no. 7 (July 1987): 600–10.

28. *The Unicode Standard, Version 4.0* (Boston: Addison-Wesley, 2003), 1.

29. Library of Congress Network Development and MARC Standards Office, "MARC 21 Concise Format for Bibliographic Data. (2004 Concise Edition). Multiscript Records," Feb. 11, 2005. Accessed Mar. 28, 2005, www.loc.gov/marc/bibliographic/ecbdmulti.html.

30. Library of Congress Network Development and MARC Standards Office, "Record Structure. Leader," *MARC 21 Specifications for Record Structure, Character Sets, and Exchange Media*, Jan. 2000. Accessed Mar. 28, 2005, www.loc.gov/marc/specifications/specrecstruc.html#leader.

31. Library of Congress Network Development and MARC Standards Office, "MARC 21 Concise Format for Bibliographic Data. (2004 Concise Edition). 880—Alternate Graphic Representation," Feb. 11, 2005. Accessed Mar. 28, 2005, www.loc.gov/marc/bibliographic/ecbdhold.html#mrcb880.

32. Gary Smith, personal communication, Feb. 24, 2005.

33. Library of Congress Network Development and MARC Standards Office, "Character Sets. Introduction," *MARC 21 Specifications for Record Structure, Character Sets, and Exchange Media*, Jan. 2000. Accessed Mar. 28, 2005, www.loc.gov/marc/specifications/speccharintro.html.

34. Library of Congress. Program for Cooperative Cataloging, "BIBCO Core Record Standards. 9. Guidelines for Multiple Character Sets," Jan. 18, 2005. Accessed Mar. 28, 2005, www.loc.gov/catdir/pcc/bibco/coreintro.html#9.

35. Library of Congress Network Development and MARC Standards Office, "Character Sets: Part 3— Code Table Greek Symbols," *MARC 21 Specifications for Record Structure, Character Sets, and Exchange Media*, Sep. 2004. Accessed Mar. 28, 2005, http://lcweb2.loc.gov/cocoon/codetables/2.html.

36. International Organization for Standardization, *Extension of the Latin Alphabet Coded Character Set for Bibliographic Information Interchange* (Geneva: International Organization for Standardization, 1983). ISO 5426:1983; International Organization for Standardization, *Documentation—African Coded Character Set for Bibliographic Information Interchange* (Geneva: International Organization for Standardization, 1983). ISO 6438:1983.

37. Unicode Consortium, "Unicode Standard Annex #24: Script Names," Apr. 17, 2003. Accessed Mar. 28, 2005, www.unicode.org/reports/tr24/.