APPENDIX C: CHARACTER SETS

D.1 Introduction

UNIMARC records may be encoded using either 7-bit or 8-bit character code values. The specifications for identifying and using various character sets are described in the following sections of this appendix; they are in conformance with those contained in ISO 2022. That standard should also be consulted.

UNIMARC records may also be encoded using 16-bit character code values. See D.6 ISO 10646 character set.

D.2 Framework

A matrix for all character codes possible with 7-bits is constructed as illustrated. Bits 7-5 are represented by the columns, and bits 4-1 by the rows. The ISO method of numbering is used, e.g. 7/15 not 7F for DEL.

7 bit Code Matrix

	columns									
rows	0	1			2	3	4	5	6	7
0					SP					
1										
2	32				94 graphic characters					
	control									
.	functions									
.										
.										
.										
15										DEL

A 7-bit code set accommodates 32 control functions, 94 graphic characters, SPACE, and DELETE. The individual characters are commonly referred to by their column and row position in the matrix using the notation "c/r", thus the SPACE character is 2/0. Code values are assigned according to the following rules. The first two columns of a code matrix are reserved for system control functions; columns 2-7 are for graphic characters. The two corner codes of the graphic columns are reserved for SPACE and DELETE characters.

Data may also be encoded using 8-bits per character, in which case the number of possible codes doubles, hence the code matrix doubles. Bits 8-5 are represented by the column and bits 4-1 by the rows. The 8-bit matrix has four parts which are specified for control functions and graphic characters as illustrated.

8-bit Code Matrix

		00 01		02	03	04	05	06	07		08 09		10	11	12	13	14	15
0				SP														
1																		
2		32		94 graphic characters							32		94 graphic characters					
.		control									control							
.		functions									functions							
.																		
.																		
.																		
15									DEL									

The additional bit is the left-most bit and it is 0 for a left-hand part and 1 for a right-hand part. Graphic sets may be represented by either one 7 or 8 bit combination per character or, where there are a large number of characters in the set, by multiple 7 or 8 bit combinations per character.

Use of code sets require first the designation of the sets, then the invocation of a designated set as the working set. For both 7-bit and 8-bit codes, two sets of control functions and four graphic character sets may be designated at any given time. These designated sets are called the C0, C1 and G0, G1, G2, G3 sets. In 7-bits, two C*n* sets and one G*n* set may have invoked, working set status at a given time. In 8-bits, two C*n* and two G*n* sets may be in an invoked, working set, status at a given time. The following appendix sections specify the designation and invocation of code sets in UNIMARC.

D.3 Control Function Sets

The C0 and C1 control function sets are fixed for UNIMARC. Thus they do not need to be designated and invoked in the record.

The C0 set is the set of 32 control functions defined in ISO 646. This set contains the basic transmission controls and the subfield delimiter, field terminator, and record terminator.

The C1 set is the set of control functions defined in ISO 6630, Bibliographic Control Characters. Only the NSB "Non-sorting character(s) beginning", NSE "Non-sorting character(s) ending", PLD "Partial Line Down" and PLU "Partial Line Up" functions from that set are currently allowed in UNIMARC.

In the 7-bit and 8-bit environment, the C0 set occupies columns 0 and 1 at all times. In a 7-bit record, the characters from the C1 set are represented by the two character "ESC F" where ESC is the 1/11 control function in the C0 set and F is a bit combination from columns 4 and 5. The F bit combinations associated with each of the functions defined in ISO 6630 were assigned by ISO at the time of registration of the set and are identified for ISO 6630 in section D.7 of this appendix. Note especially that in the 7-bit environment the "ESC F" substitutes for the code table bit combinations of the ISO 6630 functions.

In an 8-bit record, the C1 set resides in columns 08 and 09, and the functions are represented by their code table bit combinations.

D.4 Graphic Character Sets

The G0 graphic set for UNIMARC is always ISO 646. All of the characters in the RECORD LABEL, the DIRECTORY, and the coded fields/subfields are from ISO 646, as are the field indicators and subfield codes. Thus a record always begins with ISO 646 as the working set. Up to three additional graphic sets may be designated as G1, G2 and G3 in field 100, subfield $a, character positions 28-29, Character Sets, and positions 30-33, Additional Character Sets. If no more than four sets are used in a record, the field 100 information is all that is required to designate the graphic sets. The0y can then be invoked as needed. Note that since the RECORD LABEL, DIRECTORY, and coded data fields are all coded using ISO 646, the G1, G2, and G3 designations in field 100 can be accessed before any additional graphic sets are encountered in the record.

D.4.1 7-bit Environment

In a 7-bit character record the four designated sets are invoked using the following ISO 2022 locking shifts:

Acronym	Full Name	Bit Combination(s)	Set Invoked
SI	Shift in	0/15	G0
SO	Shift out	0/14	G1
LS2	Locking shift two	ESC 6/14	G2
LS3	Locking shift three	ESC 6/15	G3

These shifts are locking, so the set invoked remains the working set until another set is specified by a shift function.

Since the record begins with the G0 (ISO 646) set as the working set, the SI shift to the G0 set will only be used when there has been an invocation of one of the other G*n* sets as the working set. The G0 (ISO 646) set must be the working set at the end of each subfield and field since the succeeding subfield codes or directory processing require ISO 646 as the working set. This shift back to the G0 (ISO 646) set should take place before the subfield delimiter or end of field mark.

In 7-bits, a non-locking invocation of single characters from the designated G2 or G3 set is also possible. The following non-locking shifts are defined by ISO 2022:

Acronym	Full Name	Bit Combinations	Set from which Single Character Invoked
SS2	Single shift two	ESC 4/14	G2
SS3	Single shift three	ESC 4/15	G3

There is no need to reinvoke the working set after the single shifts as it is automatically reinstated after one character from the G2 or G3 set.

Examples (for clarity, bit combinations are in bold)

EX 1
SO SI
50011$aEdda S**0/14**æS**0/15**mundar.$mEnglish.$1Selections.
In this record, the ISO 5426 Extended Latin set has been designated the G1 set and the single character “æ” is accessed via an invocation of that set.
EX 2
SS2
50011$aEdda S**1/11 4/14**æmundar.$mEnglish.$1Selections.
If in EX 1 ISO 5426 had been designated a G2 set, the single shift function could be used to invoke the “æ”.
EX 3
LS2 SI LS2 SI
210##$a**1/11 6/14**Москва**0/15**$c"**1/11 6/14**Правда**0/15**"$d1968
In this record, ISO 5426 has been designated the G1 set and the basic Cyrillic set has been designated the G2 set. This field contains a Cyrillic name. Shifts into the G2 set must be made at the beginning of each subfield with shifts back into the G0 set at the end of each.

D.4.2 8-bit Environment

In an 8-bit code record the four designated sets are invoked using the following ISO 2022 locking shifts:

Acronym	Full Name	Bit Combinations	SetInvoked/ Into Columns
LS0	Locking shift zero	00/15	G0/02-07
LS1	Locking shift one	00/14	G1/02-07
LS1R	Locking shift one right	ESC 7/14	G1/10-15
LS2	Locking shift two	ESC 6/14	G2/02-07
LS2R	Locking shift two right	ESC 7/13	G2/10-15
LS3	Locking shift three	ESC 6/15	G3/02-07
LS3R	Locking shift three right	ESC 7/12	G3/10-15

These shifts are locking, so the set invoked remains the working set until another set is invoked by a shift function.

Since the record begins with the G0 set (ISO 646) in columns 02-07 and the G1 set in columns 10-15, the shift functions to those sets will only be used when there has been an invocation of the G2 or G3 set into those columns. The G0 set must be the working set in columns 02-07 at the end of each subfield and each field. The shift back to the G0 set when it has been temporarily displaced should occur before the subfield delimiter or end of field mark. The G1 set designated in field 100 is considered the default set for columns 10-15; thus it should always be restored at the end of a field that has shifted another set into those columns.

In 8-bits, non-locking single shifts are not used in UNIMARC.

Examples (for clarity, bit combinations are in bold)

EX 1
50011$aEdda Sæmundar.$mEnglish.$1 Selections.
The ISO 5426 Extended Latin set has been designated the G1 set. No shift is required to use it in the 8-bit environment.
EX 2
LS2R LS1R
50011$aEdda S**1/11 7/13**æ**1/11 7/14**mundar.$mEnglish.$1Selections.
The basic Cyrillic set has been designated the G1 set and the ISO 5426 Extended Latin set has been designated the G2 set. The G2 set is invoked to columns 10-15 using the LS2R, displacing the default G1 set. Following the use of the G2 set, the G1 set is reinvoked into columns 10-15.
EX 3
LS2R LS1R
210#$a**1/11 7/13**Москва$c"Правда**1/11 7/14"**$d1968
ISO 5426 is the default G1 set and the basic Cyrillic set has been designated the G2 set. The G2 set is invoked into columns 10-15 when needed. Since the subfield code comes from the G0 set and it is still the column 02-07 working set at the end of the $a subfield, no shift need take place before the "$c". The default G1 set is restored to columns 10-15, however, at the end of the use of the Cyrillic set in this field.
EX 4
305##$aВпервые иэдано в С.петерБурге на нем. яэ. в 1770-1784 в 4-х
LS2R LS1R
частях под эаглавием "Reise durch Ru**1/11 7/13**ß**1/11 7/14**land zur Untersuchung der drey Natur-Reiche". Ч.4 на рус. яэ. не переведена
Basic Latin and Basic Cyrillic are the designated G0 and G1sets, and Extended Latin the G2 set (100 $a/26-33 = 010203##). The Basic Latin and Cyrillic characters can be accessed without change to the settings. The German "ss" character (ß) is found in the Extended Latin set, which is invoked into columns 10-15 byLS2R (ESC 7/13), temporarily displacing Basic Cyrillic. This is then restored by LS1R(ESC 7/14).

D.5 Additional Graphic sets

In some instances more than the four graphic sets designated in field 100 may be required in a UNIMARC record. Additional sets may be substituted for the sets designated in field 100 through an escape of the form 'ESC I F'. 'I', which may be one or more characters in length, indicates the G*n* designation of the set according to the following values:

Single Byte per Character	Multiple Bytes per Character	G*n* Designation
2/8 or 2/12	2/4 2/8 or 2/4 2/12	G0
2/9 or 2/13	2/4 2/9 or 2/4 2/13	G1
2/10 or 2/14	2/4 2/10 or 2/4 2/14	G2
2/11 or 2/15	2/4 2/11 or 2/4 2/15	G3

F', the Final character, indicates the graphic set being designated. It is a bit combination from columns 4 to 7 that is assigned by ISO when the set is registered. The Final characters for the sets approved for use with UNIMARC are listed below. Final characters for other approved sets have not yet been assigned.

F	Graphic Set
4/0	ISO 646 (IRV), Basic Latin set
5/0	ISO 5426-1980, Extended Latin set
4/14	ISO Registration #37, Basic Cyrillic
5/1	ISO 5427-1984, Extended Cyrillic set

5/3	ISO 5428-1980, Greek set
4/13	ISO 6438-1983, African coded character set

If a fifth, etc., graphic set is needed in a UNIMARC field, it must first be designated through the escape sequence, then it may be invoked with shift functions as specified in Section D.4. When an additional set has been designated and invoked in a field, before the end of the field the original set specified in field 100 should be redesignated for the G*n* via an escape sequence. When a field is exited, the G0, G1, G2, G3 designated sets must be those specified in field 100.

Example (for clarity, bit combinations are alternately bold and italic)

Designation of Greek set as G1 LS1R
454 #0$1700#0$aXenophon.150010a1/11 **2/9 5/3** 1/11 7/14'Άπομνημονευματα1/11 **2/9 5/0** 1/11 7/14
Redesignation of Extended Latin set as G1 set LS1R

The record is for a Bulgarian translation of a Greek work and the language of cataloguing is English. The agency has designated in field 100 the following sets:

G0	ISO 646, Basic Latin
G1	ISO 5426, Extended Latin
G2	ISO Registration #37, Basic Cyrillic
G3	ISO DIS 5427, Extended Cyrillic

When the Greek set is needed in the 454 field to give the original title in Greek, it is designated as the G1 set via the sequence ESC 2/9 5/3 and then invoked into columns 10-15 via the sequence ESC 7/14. Before exiting the field, the Extended Latin set is restored to the G1 designation via ESC 2/9 5/0 and it is reinvoked into columns 10-15 via ESC 7/14.

D.6 ISO 10646 character set

ISO 10646, being a 16-bit character set, contains all necessary characters. This will be used for the C0, C1 and all G sets.

D.7 Character set tables

Sections D.8 through D.10 contain the code tables for some of the character sets specified for use in UNIMARC records. These character sets are reproduced with the permission of the International Organization for Standardization (ISO). Copies of the complete standards can be obtained from the ISO Central Secretariat, Case postale 56, 1211 GENEVA 20, Switzerland, and from any ISO Member Body.

D.8 Basic Control Set – ISO 646 (IRV)

This control set is the C0 set for UNIMARC records.

The following positions are the only ones to be used in UNIMARC

Position	Acronym	Name
0/14	SO	Shift Out
0/15	SI	Shift In
1/11	ESC	Escape
1/13	IS3	Information Separator Three
1/14	IS2	Information Separator Two
1/15	IS1	Information Separator One

In this Manual, the symbols for the Information Separators are :

IS1	$	(Subfield deliminator)

IS2	@	(Field separator) In most examples the end of field mark is not shown
IS3	%	(Record terminator)

D.9 Bibliographic Control Set – ISO 6630: 1986

This control set contains control functions required for filing, sorting, permuting, etc. It is the C1 set for UNIMARC records.

The following positions are the only ones to be used in UNIMARC:

Position	Acronym	Name
08/08	NSB	Non-Sorting Character(s), Beginning
08/09	NSE	Non-Sorting Character(s), End
08/11	PLD	Partial Line Down
08/12	PLU	Partial Line Up

In this Manual, the symbols for the non-sorting characters are:

NSB ≠NSB≠

NSE ≠NSE≠

PLU is used both to produce superscript text and to restore to the previous position subscript text created by the use of PLD. The reverse is also true, as is shown in the following example:

2^3+3^2 is expressed as 2≠PLU≠3≠PLD≠+3≠PLU≠2≠PLD≠

D.10 Basic Latin Set – ISO 646 (IRV)

This graphic set is specified in ISO 646. It is the default G0 set for UNIMARC records.

Position	Name	Position	Name
2/0	Space, Blank	5/0	Capital Letter P
2/1	Exclamation Mark	5/1	Capital Letter Q
2/2	Quotation Mark	5/2	Capital Letter R
2/3	Number Sign	5/3	Capital Letter S
2/4	Dollar Sign	5/4	Capital Letter T
2/5	Per Cent Sign	5/5	Capital Letter U
2/6	Ampersand	5/6	Capital Letter V
2/7	Apostrophe	5/7	Capital Letter W
2/8	Left Parenthesis	5/8	Capital Letter X
2/9	Right Parenthesis	5/9	Capital Letter Y
2/10	Asterisk	5/10	Capital Letter Z
2/11	Plus Sign	5/11	Left Square Bracket
2/12	Comma	5/12	Reverse Solidus
2/13	Hyphen, Minus Sign	5/13	Right Square Bracket
2/14	Full Stop, Period	5/14	Circumflex Accent
2/15	Solidus	5/15	Underline
3/0	Digit Zero	6/0	Grave Accent
3/1	Digit One	6/1	Small Letter a
3/2	Digit Two	6/2	Small Letter b
3/3	Digit Three	6/3	Small Letter c
3/4	Digit Four	6/4	Small Letter d
3/5	Digit Five	6/5	Small Letter e
3/6	Digit Six	6/6	Small Letter f
3/7	Digit Seven	6/7	Small Letter g
3/8	Digit Eight	6/8	Small Letter h

3/9	Digit Nine	6/9	Small Letter i
3/10	Colon	6/10	Small Letter j
3/11	Semi-colon	6/11	Small Letter k
3/12	Less than Sign	6/12	Small Letter l
3/13	Equals Sign	6/13	Small Letter m
3/14	Greater than Sign	6/14	Small Letter n
3/15	Question Mark	6/15	Small Letter o
4/0	Commercial At	7/0	Small Letter p
4/1	Capital Letter A	7/1	Small Letter q
4/2	Capital Letter B	7/2	Small Letter r
4/3	Capital Letter C	7/3	Small Letter s
4/4	Capital Letter D	7/4	Small Letter t
4/5	Capital Letter E	7/5	Small Letter u
4/6	Capital Letter F	7/6	Small Letter v
4/7	Capital Letter G	7/7	Small Letter w
4/8	Capital Letter H	7/8	Small Letter x
4/9	Capital Letter I	7/9	Small Letter y
4/10	Capital Letter J	7/10	Small Letter z
4/11	Capital Letter K	7/11	Left Curly Bracket
4/12	Capital Letter L	7/12	Vertical Line
4/13	Capital Letter M	7/13	Right Curly Bracket
4/14	Capital Letter N	7/14	Tilde
4/15	Capital Letter O		

N.B. If this set is used in combination with ISO 5426 positions 5/15, 6/0 and 7/14 in ISO 646 should not be used. Positions 5/8, 4/1 and 4/5 in ISO 5426 should be used instead.

History

2020	Previous appendix J.