

Character Sets

Chapter overview

Chapter 4 describes the various character sets used in OCLC-MARC records.

In this chapter

This chapter includes the following sections:

Section	Page
4.1 OCLC-MARC Character Set	4:2
4.2 Non-Latin Script Character Set	4:9

4.1 OCLC-MARC Character Set

Introduction

This section defines the character sets used in OCLC-MARC records. The characters are represented on the records as 8-bit binary codes. The hexadecimal representation of all 256 possible codes is followed by the definition of the corresponding character or function code.

Hexadecimal codes

- Codes 00 through 1A (except 01, 02, 03, and 11), 7F, and 1C are used only for data transmission control and should not appear in any record
- Codes 01, 02, 03, and 11 (in addition to being used for data transmission control) formerly identified the transaction type in byte 22 of a record leader
- Code 50 (defined in the table as the letter *p*) was also formerly used in byte 22 of the leader to identify an *All Produce* transaction
- Codes 90, 91, 93, and 94 are undefined in the table and were formerly used in byte 22 of the leader to identify an offline transaction

Nonstandard character sets

The nonstandard character sets are produced by means of an escape code (hexadecimal 1B) followed by a code that indicates the character set (with characters for consideration). The new character set remains in effect until another escape code indicates a change to a different character set.

For example, the character sequence $N^{-2}-2$ is represented as $NESCp-2ESC s-2$ (in hexadecimal, 4E 1B 70 **2D 32** 1B 73 **2D 32**). Note that code **2D 32** represents both the normal -2 and the superscript $-^2$ in the 2 character sets. All fields are assumed to begin in the standard character set.

Diacritics

Diacritical marks are characters printed over or under another character. They have codes in the range E0 through FF and follow the character to which they apply. For example, *ö* (lowercase o umlaut) is represented as *o¨* (in hexadecimal 6F E8).

4.1 OCLC-MARC Character Set (continued)

Standard character set

H E 2 X	1	0	1	2	3	4	5	6	7	8	9	A	B	C	D	E	F
				SP	0	@	P		p				'			?	،
			!	1	A	Q	a	q				ł	ł			˘	،
			"	2	B	R	b	r				Ø	ø			˘	.
			#	3	C	S	c	s				Đ	đ			˘	..
			\$	4	D	T	d	t				Đ	đ			˘	o
			%	5	E	U	e	u				Æ	æ			˘	=
			&	6	F	V	f	v				Œ	œ			˘	—
			'	7	G	W	g	w				'	"			˘	،
			(8	H	X	h	x				·	ı			˘	،
)	9	I	Y	i	y				ı	£			˘	،
			*	:	J	Z	j	z				®	đ			˘	،
		ESC	+	;	K	[k	{				±				˘	،
			,	<	L	\	l					Œ	œ			˘	
		RECORD TERM.	-	=	M]	m	}				Ū	ū			˘	
		FIELD TERM.	.	>	N		n					'	ł			˘	'
		DELIMITER	/	?	O		o									˘	

4.1 OCLC-MARC Character Set (continued)

Standard character set (continued)

Hex	Graphic	Name and/or Function	Hex	Graphic	Name and/or Function
00	Null	20	20	Space	
01		Start of heading	21	!	Exclamation point
02		Start of text	22		Quotation marks
03		End of text	23	#	Number sign
04		End of transmission	24	\$	Dollar sign
05		Enquiry	25	%	Percent sign
06		Acknowledge	26	&	Ampersand
07		Bell	27		Apostrophe
08		Backspace	28	(Opening parenthesis
09		Horizontal tabulation	29)	Closing parenthesis
0A		Line feed	2A	*	Asterisk
0B		Vertical tabulation	2B	+	Plus
0C		Form feed	2C	,	Comma
0D		Carriage return	2D	-	Hyphen (minus)
0E		Shift out	2E	.	Period (decimal point)
0F		Shift in	2F	/	Slash
10		Data link escape	30	0	
11		Device control 1	31	1	
12		Device control 2	32	2	
13		Device control 3	33	3	
14		Device control 4	34	4	
15		Negative acknowledge	35	5	
16		Synchronous idle	36	6	
17		End of transmission block	37	7	
18		Cancel	38	8	
19		End of medium	39	9	
1A		Substitute	3A	:	Colon
1B		Escape	3B	;	Semicolon
1C		File separator (FS)	3C	<	Less than
1D		Record terminator (GS)	3D	=	Equals
1E	¶	Field terminator (RS)	3E	>	Greater than
1F		Subfield delimiter (double dagger) (US)	3F	?	Question mark

4.1 OCLC-MARC Character Set (continued)

Standard character set (continued)

Hex	Graphic	Name and/or Function	Hex	Graphic	Name and/or Function
40	@	Commercial at sign	60	`	Spacing grave/grave accent
41	A		61	a	
42	B		62	b	
43	C		63	c	
44	D		64	d	
45	E		65	e	
46	F		66	f	
47	G		67	g	
48	H		68	h	
49	I		69	i	
4A	J		6A	j	
4B	K		6B	k	
4C	L		6C	l	
4D	M		6D	m	
4E	N		6E	n	
4F	O		6F	o	
50	P		70	p	
51	Q		71	q	
52	R		72	r	
53	S		73	s	
54	T		74	t	
55	U		75	u	
56	V		76	v	
57	W		77	w	
58	X		78	x	
59	Y		79	y	
5A	Z		7A	z	
5B	[Opening bracket	7B	{	Left curly bracket/opening curly bracket
5C	\	Reverse slash	7C		Illegal character (fill character)
5D]	Closing bracket	7D	}	Right curly bracket/closing curly bracket
5E	^	Spacing circumflex/circumflex accent	7E		Spacing tilde/tilde
5F	_	Spacing underscore/low line	7F		Delete

4.1 OCLC-MARC Character Set (continued)

Standard character set (continued)

Hex	Graphic	Name and/or Function	Hex	Graphic	Name and/or Function
80			A0		
81			A1	Ł	Polish L, uppercase
82			A2	Ø	Scandinavian O with slash, uppercase
83			A3	Ð	D with crossbar, uppercase
84			A4	Þ	Icelandic thorn, uppercase
85			A5	Æ	AE, uppercase
86			A6	Œ	OE, uppercase
87			A7	ˆ	Miägkiï znak
88			A8	·	Dot in middle of line
89			A9	♭	Musical flat
8A			AA	®	Subscript patent mark
8B			AB	±	Plus or minus
8C			AC	Ŏ	Hooked O, uppercase
8D			AD	Ū	Hooked U, uppercase
8E			AE	ʻ	Alif
8F			AF		
90			B0	ˆ	Ayn
91			B1	ł	Polish L, lowercase
92			B2	ø	Scandinavian o with slash, lowercase
93			B3	đ	d with crossbar, lowercase
94			B4	þ	Icelandic thorn, lowercase
95			B5	æ	ae, lowercase
96			B6	œ	oe, lowercase
97			B7	“	Tvërды znak
98			B8	ı	Turkish I, lowercase
99			B9	£	British pound
9A			BA	ä	Eth
9B			BB		
9C			BC	ŏ	Hooked o, lowercase
9D			BD	ū	Hooked u, lowercase
9E			BE	ℓ	Script e/, lowercase
9F			BF		

4.1 OCLC-MARC Character Set (continued)

Standard character set (continued)

Hex	Graphic	Name and/or Function	Hex	Graphic	Name and/or Function
C0	°	Degree sign	E0	?	Pseudo question mark
C1			E1	`	Grave
C2	®	Sound recording copyright sign	E2	´	Acute
C3	©	Copyright sign	E3	ˆ	Circumflex
C4	#	Musical sharp sign	E4	˜	Tilde
C5	¿	Inverted question mark	E5	˘	Macron
C6	¡	Inverted exclamation mark	E6	ˆ	Breve
C7	ß	Eszett	E7	·	Superior dot
C8	€	Euro sign	E8	¨	Umlaut or dieresis
C9			E9	ˇ	Hacek
CA			EA	°	Circle or angstrom
CB			EB	ſ	Ligature (left half)
CC			EC	˘	Ligature (right half)
CD			ED	ˆ	High comma diacritical
CE			EE	¨	Double acute
CF			EF	◌̣	Candrabindu
D0			F0	¸	Cedilla
D1			F1	◌̣	Right hook
D2			F2	·	Dot below character
D3			F3	¨	Double dot below character
D4			F4	◌̣	Circle below character
D5			F5	=	Double underscore
D6			F6	_	Underscore
D7			F7	◌̣	Left hook
D8			F8	◌̣	Right cedilla
D9			F9	◌̣	Upadhamaniya
DA			FA	ˆ	Double tilde (left half)
DB			FB	˘	Double tilde (right half)
DC			FC		
DD			FD		
DE			FE	ˆ	High comma (centered)
DF			FF		

4.1 OCLC-MARC Character Set (continued)

Subscript character set Reach by hexadecimal 1B 62 (ESCb). Return to standard character set with hexadecimal 1B 73 (ESCs).

Hex	Graphic	Name and/or Function	Hex	Graphic	Name and/or Function
28	(Open parenthesis	33	3	
29)	Close parenthesis	34	4	
2B	+	Plus	35	5	
2D	-	Minus	36	6	
30	0		37	7	
31	1		38	8	
32	2		39	9	

Superscript character set Reach by hexadecimal 1B 70 (EScP). Return to standard character set with hexadecimal 1B 73 (ESCs).

Hex	Graphic	Name and/or Function	Hex	Graphic	Name and/or Function
28	(Open parenthesis	33	3	
29)	Close parenthesis	34	4	
2B	+	Plus	35	5	
2D	-	Minus	36	6	
30	0		37	7	
31	1		38	8	
32	2		39	9	

4.2 Non-Latin Script Character Set

Scripts and languages supported

OCLC supports the following scripts in WorldCat. Records may contain more than one non-Latin script at any location, including within the same field.

Script	Examples of supported languages
Arabic	Arabic, Persian, Urdu, Azerbaijani
Bengali	Bengali, Assamese
Chinese	Chinese
Cyrillic	Russian, Bulgarian, Serbian, Ukrainian
Devanagari	Hindi, Marathi, Sanskrit, Nepali, Sherpa
Greek	Greek
Hebrew	Hebrew
Japanese	Japanese
Korean	Korean
Tamil	Tamil
Thai	Thai

Valid character sets: MARC-8

Arabic, CJK, Cyrillic, Greek, and Hebrew

Character sets for these scripts given in *MARC 21 Specifications for Record Structure, Character Sets, and Exchange Media* on the Library of Congress Web site at: <http://www.loc.gov/marc/specifications/spechome.html> define the scope of valid characters in WorldCat. The MARC-8 character set is the subset of Unicode characters approved for use in MARC 21 cataloging.

The following list defines the scope of valid characters in WorldCat for Arabic (including Persian), CJK, Cyrillic, Greek, and Hebrew scripts:

- 33(hex) [ASCII graphic: **3**] Basic Arabic
- 34(hex) [ASCII graphic: **4**] Extended Arabic
- 31(hex) [ASCII graphic: **1**] Chinese, Japanese, Korean (EACC)
- 4E(hex) [ASCII graphic: **N**] = Basic Cyrillic
- 51(hex) [ASCII graphic: **Q**] = Extended Cyrillic
- 53(hex) [ASCII graphic: **S**] = Basic Greek
- 32(hex) [ASCII graphic: **2**] = Basic Hebrew

If you receive records output in MARC-8 data format, characters that are not supported in MARC-8 are represented by numeric character references (NCR).

4.2 Non-Latin Script Character Set (continued)

Valid character sets: **Bengali, Devanagari, Tamil, and Thai**

UTF-8 Unicode

There are no MARC-8 character sets for Bengali, Devanagari, Tamil, or Thai. OCLC implemented script identification codes for these scripts based on ISO 15924 Code Lists (<http://www.unicode.org/iso15924/codelists.html>).

The following list shows the ranges of UTF-8 Unicode characters that define valid characters for these scripts in WorldCat records:

- Bengali (character range U+0980 to U+09FF)
- Devanagari (character range U+0900 to U+097F)
- Tamil (character range U+0B80 to U+0BFF)
- Thai (character range U+0E00 to U+0E7F)

If you receive records in MARC-8 data format, characters that are not supported in MARC-8 are represented by numeric character references (NCR).

Field 066

Field 066 identifies the presence of any MARC-8 character sets for non-Latin scripts in the record. Field 066 is not used: for UTF-8 character sets.

‡c Alternate graphic character set identification Subfield ‡c contains a code identifying the MARC-8 character set used in the record. The subfield is repeated for each additional character set present. The following codes display:

\$1 Chinese, Japanese, Korean present

(3 Basic Arabic present

(4 Extended Arabic present

(N Basic Cyrillic present

(Q Extended Cyrillic present

(S Extended Greek present

(2 Basic Hebrew present

4.2 Non-Latin Script Character Set (continued)

Field 066 (continued)

Note: These character sets encode language data in the script of the language. They do not encode romanized data in Latin script. The dollar sign ("\$") means the character set has multiple bytes per character. The left paragraph mark ("(") means the character set has one byte per character.

WorldCat records can contain:

- Non-Latin script data only (or multiple non-Latin scripts if needed, one script per field.)
Or
- Latin script equivalent data only
Or
- Both non-Latin and Latin scripts. For paired non-Latin and Latin fields, the non-Latin script data is stored in field 880.

Field 880

Non-Latin characters appear in field 880 (Alternate Graphic Representation). Field 880 appears in MARC records, but does not display in online WorldCat records. The data it contains appears online in the field linked by subfield †6 (Linkage).

Definition. Fully content-designated representation, in a different script, of another field in the same record. Field 880 is linked to the associated regular field by subfield †6 (Linkage). A subfield †6 in the associated field also links that field to the 880 field. The data in field 880 may be in more than one script.

When an associated field does not exist in the record, field 880 is constructed as if it did and a reserved occurrence number (00) is used to indicate the special situation.

Indicators. Indicators in field 880 have the same meaning and values as the appropriate indicators in the available associated field and are not described in this section. See the description of the specific associated field.

Subfield codes. Subfield codes in field 880 parallel those in the associated field, with the addition of **linking subfield †6**.

4.2 Non-Latin Script Character Set (continued)

Subfield †6 (continued)

When there is no associated field to which a field 880 is linked, the occurrence number in subfield †6 is 00. It is used if an agency wants to separate scripts in a record (see Multiscript Records). The linking tag part of subfield †6 will contain the tag that the associated regular field would have had if it had existed in the record.

880 ##†6530-00/(2/r†a[Additional physical form available information in Hebrew script]
[Field 880 is not linked to an associated field. The occurrence number is 00.]

The occurrence number is followed immediately by a slash (/) and the **script identification code**. This code identifies the alternate script found in the field. The following codes are used:

Code -- Script

(3 Arabic
(B Latin
(\$1 Chinese, Japanese, Korean
(N Cyrillic
(S Greek
(2 Hebrew
880 1#†6100-01/(N†a[Heading in Cyrillic script]

The entire field need not be in the script identified in subfield †6. If more than one script is present in the field, subfield †6 will contain the identification of the first alternate script encountered in a left-to-right scan of the field.

Note also that the script identification code is used in field 880, subfield †6, but this data element is not generally used for subfield †6 of the associated regular field. In the associated field, the data is assumed to be the primary script(s) for the record.

In a MARC record, the contents of field 880 are always recorded in their logical order, from the first character to the last, regardless of field orientation. For a display of the field, the default field orientation is left-to-right. When the field contains text that has a right-to-left orientation, the script identification code is followed by a slash (/) and the field orientation code. The MARC field orientation code for right-to-left scripts is the letter **r**. The orientation code is only included in fields with right-to-left orientation, since left-to-right orientation is the default orientation in 880 fields. (See MARC 21 Specifications for Record Structure, Character Sets, and Exchange Media for a detailed description of field orientation.)

110 2#†6880-01†a[Heading in Latin script]
880 2#†6110-01/(2/r†a[Heading in Hebrew script linked to associated field]

Note that the orientation code is used in field 880, subfield †6, but this data element is not generally used for subfield †6 of the associated regular field. In the associated field, the data is assumed to be the usual orientation of the primary script(s) for the record.