UTF-8 & Latex Encodings of ISO-8859 (Latin-1) Character Set

Technical Report · March 2022

DOI: 10.13140/RG.2.2.18402.61121

1 author:

Manuel José Fernández Iglesias
atlanTTic - University of Vigo
248 PUBLICATIONS **1,431** CITATIONS

SEE PROFILE

UTF-8 & LaTeX Encodings of ISO-8859-1 (Latin-1) Character Set

Manuel J. Fernández Iglesias

https://desire.webs.uvigo.gal

Abstract

The ISO-8859-1 character set, also known as Latin-1, is an 8-bit character set that includes all the characters used in Western European alphabets based on the Latin alphabet. UTF-8 is a variable-length character encoding format that became the dominant encoding for internet technologies and most computing platforms. Traditional 7-bit ASCII characters (ISO-8859-1 characters from `0x00` to `0x7F`) are encoded in UTF-8 by means of single-byte codes that match the ASCII codes. The rest of the ISO-8859-1 codes (from `0x80` to `0xfF`) are encoded using two bytes. This document discusses the UTF-8 encoding of the ISO-8859-1 set and includes the LaTeX commands necessary to obtain all the characters in it.

1 Introducción

The ISO-8859-1 (Latin-1) character set is an 8-bit or 256-character character set endorsed by the International Organization for Standardization (ISO) that includes the characters used in Western European languages based on the Latin alphabet. As its name implies, it is a subset of ISO-8859, which addresses other writing systems or alphabets such as Cyrillic, Hebrew, or Arabic. Until the popularization of the UTF-8 encoding, it was the encoding used by most Unix systems, as well as by the Microsoft Windows operating system. This character set is also known as extended ASCII because its first 128 characters are the same as the ASCII standard developed by the American Standards Association, now the American National Standards Institute (ANSI).

UTF-8 (8-bit Unicode Transformation Format) is variable-width character encoding defined by the Unicode standard and also adopted by the Internet Engineering Task Force (IETF) in RFC 2277 (BCP 18) for future internet standards work, replacing single-byte character sets such as ISO-8859-1. UTF-8 is the dominant encoding in present-day computing environments and internet technologies, accounting for 98% of all web pages, and up to 100.0% for some languages and computing scenarios.

UTF-8 directly encodes the traditional 7-bit ASCII characters (ISO-8859-1 characters from `0x00` to `0x7F`), so any ASCII message or document is rendered unchanged. UTF-8 encodes practically all commonly used symbols worldwide, such as the characters of any alphabet (e.g., Latin, Cyrillic, Chinese, Japanese, Korean, etc.) or mathematical symbols. For this, it utilizes 2, 3 or 4 bytes. In the case of the Western European symbols in ISO-8859-1, the remaining 128 glyphs (codes from `0x80` to `0xFF`) are encoded with two bytes according to the model outlined in Table 1.

This document collects the glyphs in the ISO-8859-1 characters set and details its UTF-8 encoding, as well as the LaTeX commands necessary to obtain all the mentioned glyphs. To use any of these characters directly in a LaTeX document, the selected input and output encodings have to be considered.

Table 1. UTF-8 encoding of ISO-8859-1 characters (codes `0x00` – `0xFF`).

Range	UTF-8 (binary)	Comments
`0x00`-`0x7F`	`0b0xxxxxxx`	7-bit ASCII characters.
`0x80`-`0xFF`	`0b110yyyyy 0b10xxxxxx`	Accented characters and other commonly used symbols.

- In the case of directly using UTF-8 as the input encoding (e.g., the characters in `.tex` files), which is practically the standard in any modern computing scenario, it is not necessary to specify any encoding as LaTeX is configured by default to use UTF-8. In the case of using another input encoding, such as ISO 8859–1, the needed encoding should be loaded explicitly. For example, to use ISO 8859–1 we would include `\usepackage[latin1]{inputenc}` in the preamble of the LaTeX document.
- If `pdflatex` or `latex` are used, the default output encoding will be OT1 (e.g., characters in `.dvi` or `.pdf` files), so the output encoding should be set to T1 to obtain the UTF-8 glyphs. The T1 font encoding, also known as Cork encoding, provides 256 glyph slots, including all glyphs in ISO 8859–1, and allows hyphenation for most Western European languages. In addition to missing many of the symbols in the UTF-8 repertoire, OT1 does not include accented letters, so accented letters will be constructed using the `\accent` primitive instead of the accented characters in the corresponding font. This has the consequence that words using OT1 will not be hyphenated. To load the T1 output encoding use `\usepackage[T1]{fontenc}` in the preamble of your document.

Once our LaTeX document is configured to use UTF-8 as input encoding and T1 as output font encoding, we may use any glyph from the tables below directly (e.g., by copying it from another UTF-8-encoded document). Alternatively, in those cases where we cannot directly access the original glyphs, we can use the LaTeX commands provided in the tables. Many of these commands are defined in the file `utf8enc.dfu` of the base LaTeXdistribution. In the tables below, we include LaTeXcommands for those glyphs that do not appear in the original ASCII character set or that cannot be used directly in a LaTeXdocument because they serve as special characters.

2 Encoding of the ASCII set (`0x00` a `0x7F`)

The characters in the first group (cf. Tables 4 and 5) correspond to the original 7-bit ASCII characters. These are therefore single-byte symbols whose most significant bit is 0. They are grouped into control characters, digits and punctuation marks, uppercase letters and special characters, and lowercase letters and special characters, as collected in Table 2.

Table 2. ISO-8859-1/ASCII character grouping (`0x00` – `0x7F`).

Encoding	Group
`0b000X XXXX`	Control characters.
`0b001X XXXX`	Digits and punctuation marks.
`0b010X XXXX`	Uppercase letters and special characters.
`0b011X XXXX`	Lowercase letters and special characters.

The symbols listed in Table 4 for control characters are those included in the `ascii` LaTeX package, which provides the glyphs and commands to access the symbols in the *IBM PC Code Page 437 C0 Graphics*. We have included them here because such glyphs are commonly used to represent ASCII

control characters in programming environments and interface documentation. To make them available, the command `\usepackage{ascii}` must be included in the preamble. This package requires the T1 encoding above.

3 Encoding of the extended set (`0x80` a `0xFF`)

The symbols in the second group (cf. Tables 6 and 7) used to be defined as extended ASCII characters. They were encoded in one byte with the most significant bit set to 1 (i.e., codes from `0x80` to `0xFF`), and are encoded in UTF-8 using two bytes. The UTF-8 2-byte encoding is performed as follows:

$$\texttt{0b0000 0000 XXXX YYYY} \rightarrow \texttt{0b1100 00XX 10XX YYYY}$$

For example, letter "ñ" with ISO 8859–1 code `0xF1` is encoded as (cf. Table 7 in page 7):

$$\texttt{0b0000 0000 1111 0001} \rightarrow \texttt{0b1100 0011 1011 0001}$$

That is, as `0xC3B1`. The reason for this encoding is to guarantee a relevant property of the UTF-8 encoding system, namely being strictly non overlapping and self-synchronizing. The UTF-8 encoding is constructed in a way that character boundaries are easily identified by scanning for well-defined bit patterns in either direction. For example, it is not possible to confuse any one-byte UTF-8 symbol with the first or second byte of a two-byte symbol. Besides, in a transmission of UTF-8 symbols it is possible to determine the start of each symbol without restarting the transmission and byte-oriented string-searching algorithms can be used directly.

The distribution of codes and glyphs in Tables 6 and 7 is outlined in Table 3. Note that uppercase and lowercase letters differ in a single bit, the same as in the case of ASCII characters in Table 5.

Table 3. ISO-8859-1/UTF-8 character grouping (`0x80` – `0xFF`).

Encoding	Group
`0xC2 0b100X XXXX`	C1 Controls (ISO-8859-1) and symbols (UTF-8).
`0xC2 0b101X XXXX`	Symbols
`0xC3 0b100X XXXX`	Additional uppercase letters.
`0xC3 0b101X XXXX`	Additional lowercase letters.

Table 4. UTF-8 encodings from `0x00` to `0x3F`. Control characters, digits, and punctuation marks. The original IBM PC CP437 code page utilized control characters from (soh) to (us) to define the printable characters in the table. The character corresponding to (nul) was not part of the original CP437 code page.

Dec	Hex	S	Ctl	LaTeX	Dec	Hex	S	LaTeX
0	0x00	␀	(nul)	\NUL	32	0x20	␣	\textvisiblespace
1	0x01	☺	(soh)	\SOH	33	0x21	!	
2	0x02	☻	(stx)	\STX	34	0x22	"	
3	0x03	♥	(etx)	\ETX	35	0x23	#	\#
4	0x04	♦	(eot)	\EOT	36	0x24	$	\$
5	0x05	♣	(enq)	\ENQ	37	0x25	%	\%
6	0x06	♠	(ack)	\ACK	38	0x26	&	\&
7	0x07	•	(bel)	\BEL	39	0x27	'	\textquotesingle
8	0x08	◘	(bs)	\BS	40	0x28	(	
9	0x09		(tab)		41	0x29	)	
10	0x0A	◙	(lf)	\LF	42	0x2A	*	
11	0x0B	♂	(vt)	\VT	43	0x2B	+	
12	0x0C		(ff)		44	0x2C	’	\textquoteright
13	0x0D	♪	(cr)	\CR	45	0x2D	-	
14	0x0E	♫	(so)	\SO	46	0x2E	.	
15	0x0F	☼	(si)	\SI	47	0x2F	/	
16	0x10	►	(dle)	\DLE	48	0x30	0	
17	0x11	◄	(dc1)	\DCa	49	0x31	1	
18	0x12	↕	(dc2)	\DCb	50	0x32	2	
19	0x13	‼	(dc3)	\DCc	51	0x33	3	
20	0x14	¶	(dc4)	\DCd	52	0x34	4	
21	0x15	§	(nak)	\NAK	53	0x35	5	
22	0x16	▬	(syn)	\SYN	54	0x36	6	
23	0x17	↨	(etb)	\ETB	55	0x37	7	
24	0x18	↑	(can)	\CAN	56	0x38	8	
25	0x19	↓	(em)	\EM	57	0x39	9	
26	0x1A		(eof)		58	0x3A	:	
27	0x1B	←	(esc)	\ESC	59	0x3B	;	
28	0x1C	└	(fs)	\FS	60	0x3C	<	
29	0x1D	↔	(gs)	\GS	61	0x3D	=	
30	0x1E	▲	(rs)	\RS	62	0x3E	>	
31	0x1F	▼	(us)	\US	63	0x3F	?	

Table 5. UTF-8 encodings from `0x40` to `0x7F`. Uppercase letters, lowercase letters and special characters. Command `\char` returns the glyph corresponding to the numeric code passed as an argument. On the other side, TEX operator `‘` (left quote) returns the character code of a glyph regardless of whether it has a special meaning in LATEX. Thus, the composition `\char‘C` returns the glyph corresponding to character `C`. `\DEL` is provided by the `ascii` package.

Dec	Hex	S	LATEX	Dec	Hex	S	LATEX
64	0x40	@		96	0x60	‘	\textquoteleft
65	0x41	A		97	0x61	a	
66	0x42	B		98	0x62	b	
67	0x43	C		99	0x63	c	
68	0x44	D		100	0x64	d	
69	0x45	E		101	0x65	e	
70	0x46	F		102	0x66	f	
71	0x47	G		103	0x67	g	
72	0x48	H		104	0x68	h	
73	0x49	I		105	0x69	i	
74	0x4A	J		106	0x6A	j	
75	0x4B	K		107	0x6B	k	
76	0x4C	L		108	0x6C	l	
77	0x4D	M		109	0x6D	m	
78	0x4E	N		110	0x6E	n	
79	0x4F	O		111	0x6F	o	
80	0x50	P		112	0x70	p	
81	0x51	Q		113	0x71	q	
82	0x52	R		114	0x72	r	
83	0x53	S		115	0x73	s	
84	0x54	T		116	0x74	t	
85	0x55	U		117	0x75	u	
86	0x56	V		118	0x76	v	
87	0x57	W		119	0x77	w	
88	0x58	X		120	0x78	x	
89	0x59	Y		121	0x79	y	
90	0x5A	Z		122	0x7A	z	
91	0x5B	[		123	0x7B	{	\char‘\{
92	0x5C	\	\char‘\\	124	0x7C	|	
93	0x5D	]		125	0x7D	}	\char‘\}
94	0x5E	ˆ	\^{}	126	0x7E	˜	\~{}
95	0x5F	_	\char‘_	127	0x7F	⌂	\DEL

Table 6. UTF-8 encodings from `0xC280` to `0xC2BF`. Miscellaneous symbols. The first *extended* 32 ISO 8859–1 codes correspond to non-printable control characters known as C1 Controls. Column S identifies the encoded UTF-8 glyphs in this case. For the last 32 characters, Column LaTeXlists the commands to get the corresponding symbols as defined in LaTeX distribution file `utf8enc.dfu`.

ISO	UTF-8	Ctl	S	LaTeX	ISO	UTF-8	S	LaTeX
0x80	0xC280	(pad)	€	\texteuro	0xA0	0xC2A0		\nobreakspace
0x81	0xC281	(hop)			0xA1	0xC2A1	¡	\textexclamdown, !‘
0x82	0xC282	(bph)	‚	\quotesinglbase	0xA2	0xC2A2	¢	\textcent
0x83	0xC283	(nbh)	ƒ	\textit{f}	0xA3	0xC2A3	£	\textsterling, \pounds
0x84	0xC284	(ind)	„	\quotedblbase	0xA4	0xC2A4	¤	\textcurrency
0x85	0xC285	(nel)	…	\dots	0xA5	0xC2A5	¥	\textyen
0x86	0xC286	(ssa)	†	\dag	0xA6	0xC2A6	¦	\textbrokenbar
0x87	0xC287	(esa)	‡	\ddag	0xA7	0xC2A7	§	\textsection, \S
0x88	0xC288	(hts	^	\textasciicircum	0xA8	0xC2A8	¨	\textasciidieresis
0x89	0xC289	(htj)	‰	\textperthousand	0xA9	0xC2A9	©	\textcopyright
0x8A	0xC28A	(lts)	Š	\v{S}	0xAA	0xC2AA	a	\textordfeminine
0x8B	0xC28B	(pld)	‹	\guilsinglleft	0xAB	0xC2AB	«	\guillemotleft
0x8C	0xC28C	(plu)	Œ	\OE	0xAC	0xC2AC	¬	\textlnot
0x8D	0xC28D	(ri)			0xAD	0xC2AD		\-
0x8E	0xC28E	(ss2)	Ž	\v{Z}	0xAE	0xC2AE	®	\textregistered
0x8F	0xC28F	(ss3)			0xAF	0xC2AF	¯	\textasciimacron
0x90	0xC290	(dcs)			0xB0	0xC2B0	°	\textdegree
0x91	0xC291	(pu1)	‘		0xB1	0xC2B1	±	\textpm
0x92	0xC292	(pu2)	’		0xB2	0xC2B2	²	\texttwosuperior
0x93	0xC293	(sts)	“	‘‘	0xB3	0xC2B3	³	\textthreesuperior
0x94	0xC294	(cch)	”	’’	0xB4	0xC2B4	´	\textasciiacute
0x95	0xC295	(mw)	•	\textbullet	0xB5	0xC2B5	µ	\textmu
0x96	0xC296	(spa)	–		0xB6	0xC2B6	¶	\textparagraph, \P
0x97	0xC297	(epa)	—	--	0xB7	0xC2B7	·	\textperiodcentered
0x98	0xC298	(sos)	~	\textasciitilde	0xB8	0xC2B8	¸	\c{}, \c\
0x99	0xC299	(sgci)	™	\texttrademark	0xB9	0xC2B9	¹	\textonesuperior
0x9A	0xC29A	(sci)	š	\v{s}	0xBA	0xC2BA	º	\textordmasculine
0x9B	0xC29B	(csi)	›	\guilsinglright	0xBB	0xC2BB	»	\guillemotright
0x9C	0xC29C	(st)	œ	\oe	0xBC	0xC2BC	¼	\textonequarter
0x9D	0xC29D	(osc)			0xBD	0xC2BD	½	\textonehalf
0x9E	0xC29E	(pm)	ž	\v{z}	0xBE	0xC2BE	¾	\textthreequarters
0x9F	0xC29F	(apc)	Ÿ	\"{Y}	0xBF	0xC2BF	¿	\textquestiondown, ?‘

Table 7. UTF-8 encodings from `0xC380` to `0xC3BF`. Accented letters in the Latin-1 supplement. We can see that the uppercase and lowercase letters differ in a single bit, the same as in the case of ASCII characters from block `0x40` to `0x7F`.

ISO	UTF-8	Char	LaTeX	ISO	UTF-8	Char	LaTeX
0xC0	0xC380	À	\‘{A}	0xE0	0xC3A0	à	\‘{a}
0xC1	0xC381	Á	\’{A}	0xE1	0xC3A1	á	\’{a}
0xC2	0xC382	Â	\^{A}	0xE2	0xC3A2	â	\^{a}
0xC3	0xC383	Ã	\~{A}	0xE3	0xC3A3	ã	\~{a}
0xC4	0xC384	Ä	\"{A}	0xE4	0xC3A4	ä	\"{a}
0xC5	0xC385	Å	\AA	0xE5	0xC3A5	å	\aa
0xC6	0xC386	Æ	\AE	0xE6	0xC3A6	æ	\ae
0xC7	0xC387	Ç	\c{C}	0xE7	0xC3A7	ç	\c{c}
0xC8	0xC388	È	\‘{E}	0xE8	0xC3A8	è	\‘{e}
0xC9	0xC389	É	\’{E}	0xE9	0xC3A9	é	\’{e}
0xCA	0xC38A	Ê	\^{E}	0xEA	0xC3AA	ê	\^{e}
0xCB	0xC38B	Ë	\"{E}	0xEB	0xC3AB	ë	\"{e}
0xCC	0xC38C	Ì	\‘{I}	0xEC	0xC3AC	ì	\‘{i}
0xCD	0xC38D	Í	\’{I}	0xED	0xC3AD	í	\’{i}
0xCE	0xC38E	Î	\^{I}	0xEE	0xC3AE	î	\^{i}
0xCF	0xC38F	Ï	\"{I}	0xEF	0xC3AF	ï	\"{i}
0xD0	0xC390	Đ	\DH	0xF0	0xC3B0	ð	\dh
0xD1	0xC391	Ñ	\~{N}	0xF1	0xC3B1	ñ	\~{n}
0xD2	0xC392	Ò	\‘{O}	0xF2	0xC3B2	ò	\‘{o}
0xD3	0xC393	Ó	\’{O}	0xF3	0xC3B3	ó	\’{o}
0xD4	0xC394	Ô	\^{O}	0xF4	0xC3B4	ô	\^{o}
0xD5	0xC395	Õ	\~{O}	0xF5	0xC3B5	õ	\~{o}
0xD6	0xC396	Ö	\"{O}	0xF6	0xC3B6	ö	\"{o}
0xD7	0xC397	×	\texttimes	0xF7	0xC3B7	÷	\textdiv
0xD8	0xC398	Ø	\O	0xF8	0xC3B8	ø	\o
0xD9	0xC399	Ù	\‘{U}	0xF9	0xC3B9	ù	\‘{u}
0xDA	0xC39A	Ú	\’{U}	0xFA	0xC3BA	ú	\’{u}
0xDB	0xC39B	Û	\^{U}	0xFB	0xC3BB	û	\^{u}
0xDC	0xC39C	Ü	\"{U}	0xFC	0xC3BC	ü	\"{u}
0xDD	0xC39D	Ý	\’{Y}	0xFD	0xC3BD	ý	\’{y}
0xDE	0xC39E	Þ	\TH	0xFE	0xC3BE	þ	\th
0xDF	0xC39F	ß	\ss	0xFF	0xC3BF	ÿ	\"{y}