Character Sets and Encodings

Fall Term 2023-24

H. Wenger & C. Grothoff | BFH-TI

Agenda

- **Background**
- **Typography**
- **History**
- **Unicode**
- **UTF-8**

Background

Objectives

How do computers represent characters?

- Glyphs, ligatures and fonts
- Historical standards
- Unicode 5 layer architecture:
  1. Abstract Character Repertoire
  2. Coded Character Set
  3. Character Encoding Form
  4. Character Encoding Scheme
  5. Transfer Encoding Syntax

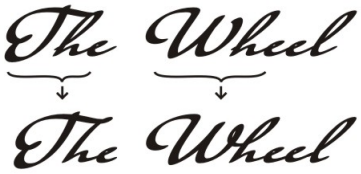
Typography

Glyph

“A grapheme (computing: character) is the smallest unit of a writing system of any given language.” –https://en.wikipedia.org/wiki/Grapheme

“A glyph (computing: shape) is an elemental symbol within an agreed set of symbols, intended to represent a readable character for the purposes of writing.” –https://en.wikipedia.org/wiki/Glyph

Ligature



A ligature is the combination of two or more graphemes (or letters) into a single glyph.

Font

A font *defines size, weight and style of a* **typeface**.

A typeface *is a set of glyphs that share common design features.*

A font may not define a unique glyph for every character. For example, the Latin and Greek ’A’ may be different characters sharing the same glyph.

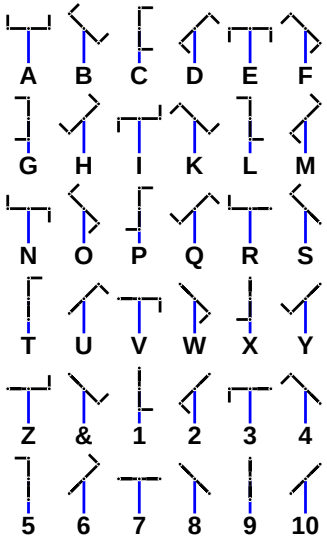
Character

Unit of information

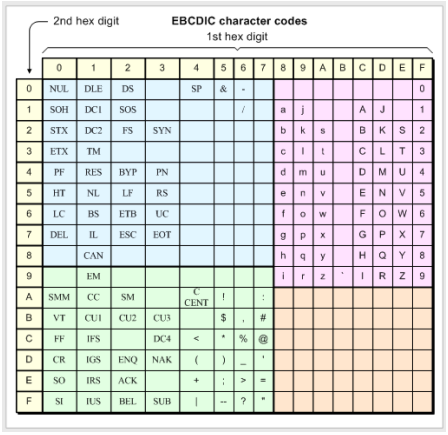
Used for organization, control or representation of textual data

History

Telegraph (Chappe, 1792)



Extended Binary Coded Decimal Interchange Code (1963)



EBCDIC character codes

2nd hex digit \ 1st hex digit	0	1	2	3	4	5	6	7	8	9	A	B	C	D	E	F	
0	NUL	DLE	DS		SP	&	-										0
1	SOH	DC1	SOS				/		a	j			A	J			1
2	STX	DC2	FS	SYN					b	k	s		B	K	S		2
3	ETX	TM							c	l	t		C	L	T		3
4	PF	RES	BYP	PN					d	m	u		D	M	U		4
5	HT	NL	LF	RS					e	n	v		E	N	V		5
6	LC	BS	ETB	UC					f	o	w		F	O	W		6
7	DEL	IL	ESC	EOT					g	p	x		G	P	X		7
8		CAN							h	q	y		H	Q	Y		8
9		EM							i	r	z	`	I	R	Z		9
A	SMM	CC	SM		¢ CENT	!		¬									
B	VT	CU1	CU2	CU3		$	,	#									
C	FF	IFS		DC4	<	*	%	@									
D	CR	IGS	ENQ	NAK	(	)	_	'									
E	SO	IRS	ACK		+	;	>	=									
F	SI	IUS	BEL	SUB	|	¬	?	"									

US ASCII (1963)

ASCII TABLE

Decimal	Hexadecimal	Binary	Octal	Char
0	0	0	0	*[NULL]*
1	1	1	1	*[START OF HEADING]*
2	2	10	2	*[START OF TEXT]*
3	3	11	3	*[END OF TEXT]*
4	4	100	4	*[END OF TRANSMISSION]*
5	5	101	5	*[ENQUIRY]*
6	6	110	6	*[ACKNOWLEDGE]*
7	7	111	7	*[BELL]*
8	8	1000	10	*[BACKSPACE]*
9	9	1001	11	*[HORIZONTAL TAB]*
10	A	1010	12	*[LINE FEED]*
11	B	1011	13	*[VERTICAL TAB]*
12	C	1100	14	*[FORM FEED]*
13	D	1101	15	*[CARRIAGE RETURN]*
14	E	1110	16	*[SHIFT OUT]*
15	F	1111	17	*[SHIFT IN]*
16	10	10000	20	*[DATA LINK ESCAPE]*
17	11	10001	21	*[DEVICE CONTROL 1]*
18	12	10010	22	*[DEVICE CONTROL 2]*
19	13	10011	23	*[DEVICE CONTROL 3]*
20	14	10100	24	*[DEVICE CONTROL 4]*
21	15	10101	25	*[NEGATIVE ACKNOWLEDGE]*
22	16	10110	26	*[SYNCHRONOUS IDLE]*
23	17	10111	27	*[ENG OF TRANS. BLOCK]*
24	18	11000	30	*[CANCEL]*
25	19	11001	31	*[END OF MEDIUM]*
26	1A	11010	32	*[SUBSTITUTE]*
27	1B	11011	33	*[ESCAPE]*
28	1C	11100	34	*[FILE SEPARATOR]*
29	1D	11101	35	*[GROUP SEPARATOR]*
30	1E	11110	36	*[RECORD SEPARATOR]*
31	1F	11111	37	*[UNIT SEPARATOR]*
32	20	100000	40	*[SPACE]*
33	21	100001	41	!
34	22	100010	42	"
35	23	100011	43	#
36	24	100100	44	$
37	25	100101	45	%
38	26	100110	46	&
39	27	100111	47	'
40	28	101000	50	(
41	29	101001	51	)
42	2A	101010	52	*
43	2B	101011	53	+
44	2C	101100	54	,
45	2D	101101	55	-
46	2E	101110	56	.
47	2F	101111	57	/
48	30	110000	60	**0**
49	31	110001	61	**1**
50	32	110010	62	**2**
51	33	110011	63	**3**
52	34	110100	64	**4**
53	35	110101	65	**5**
54	36	110110	66	**6**
55	37	110111	67	**7**
56	38	111000	70	**8**
57	39	111001	71	**9**
58	3A	111010	72	:
59	3B	111011	73	;
60	3C	111100	74	<
61	3D	111101	75	=
62	3E	111110	76	>
63	3F	111111	77	?
64	40	1000000	100	@
65	41	1000001	101	**A**
66	42	1000010	102	**B**
67	43	1000011	103	**C**
68	44	1000100	104	**D**
69	45	1000101	105	**E**
70	46	1000110	106	**F**
71	47	1000111	107	**G**
72	48	1001000	110	**H**
73	49	1001001	111	**I**
74	4A	1001010	112	**J**
75	4B	1001011	113	**K**
76	4C	1001100	114	**L**
77	4D	1001101	115	**M**
78	4E	1001110	116	**N**
79	4F	1001111	117	**O**
80	50	1010000	120	**P**
81	51	1010001	121	**Q**
82	52	1010010	122	**R**
83	53	1010011	123	**S**
84	54	1010100	124	**T**
85	55	1010101	125	**U**
86	56	1010110	126	**V**
87	57	1010111	127	**W**
88	58	1011000	130	**X**
89	59	1011001	131	**Y**
90	5A	1011010	132	**Z**
91	5B	1011011	133	**[**
92	5C	1011100	134	\
93	5D	1011101	135	**]**
94	5E	1011110	136	^
95	5F	1011111	137	_
96	60	1100000	140	
97	61	1100001	141	**a**
98	62	1100010	142	**b**
99	63	1100011	143	**c**
100	64	1100100	144	**d**
101	65	1100101	145	**e**
102	66	1100110	146	**f**
103	67	1100111	147	**g**
104	68	1101000	150	**h**
105	69	1101001	151	**i**
106	6A	1101010	152	**j**
107	6B	1101011	153	**k**
108	6C	1101100	154	**l**
109	6D	1101101	155	**m**
110	6E	1101110	156	**n**
111	6F	1101111	157	**o**
112	70	1110000	160	**p**
113	71	1110001	161	**q**
114	72	1110010	162	**r**
115	73	1110011	163	**s**
116	74	1110100	164	**t**
117	75	1110101	165	**u**
118	76	1110110	166	**v**
119	77	1110111	167	**w**
120	78	1111000	170	**x**
121	79	1111001	171	**y**
122	7A	1111010	172	**z**
123	7B	1111011	173	**{**
124	7C	1111100	174	|
125	7D	1111101	175	**}**
126	7E	1111110	176	~
127	7F	1111111	177	*[DEL]*

Control codes

Character sets often contain control codes: code positions that are not mapped to a visible character but used for output control. Examples:

CR	Carriage return
LF	Line feed
HT	Horizontal tab
CTRL-C	Interrupt
CTRL-D	End of input

End of line on Mac is `CR`, on Unix `LF` and on Windows `CR LF`.

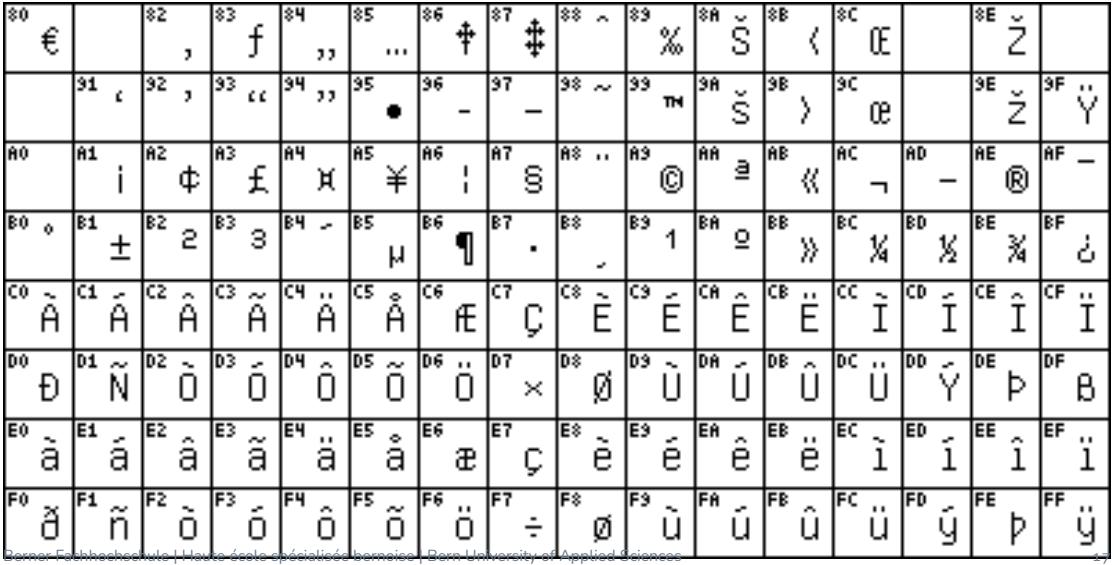
Control


AND THAT'S NOT EVEN THE WORST PART! THE WORST PART IS THAT—
U+202e
...NEVE T'NDID YEHT—
?LLEH EHT TAHW...
...UOY DID WOH
.ELOHSSA...

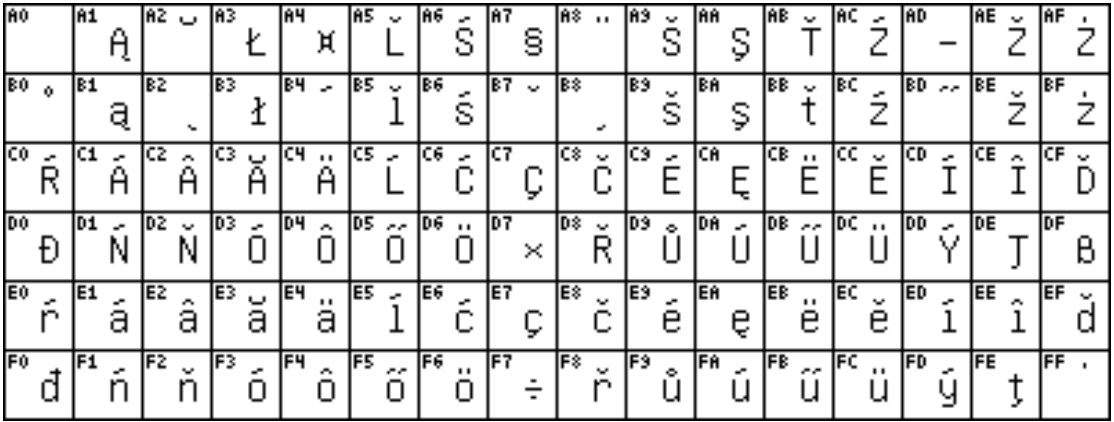
ISO 8859-1: Extended ASCII (latin1)

	A1 ¡	A2 ¢	A3 £		A5 ¥		A7 §	A8 ¤	A9 ©	AA ª	AB «				
B0 °	B1 ±	B2 ²	B3 ³		B5 µ	B6 ¶	B7 ·		B9 ¹	BA º	BB »	BC ¼	BD ½		BF ¿
C0 À	C1 Á	C2 Â	C3 Ã	C4 Ä	C5 Å	C6 Æ	C7 Ç	C8 È	C9 É	CA Ê	CB Ë	CC Ì	CD Í	CE Î	CF Ï
	D1 Ñ	D2 Ò	D3 Ó	D4 Ô	D5 Õ	D6 Ö	D7 Œ	D8 Ø	D9 Ù	DA Ú	DB Û	DC Ü	DD Ý		DF ß
E0 à	E1 á	E2 â	E3 ã	E4 ä	E5 å	E6 æ	E7 ç	E8 è	E9 é	EA ê	EB ë	EC ì	ED í	EE î	EF ï
	F1 ñ	F2 ò	F3 ó	F4 ô	F5 õ	F6 ö	F7 œ	F8 ø	F9 ù	FA ú	FB û	FC ü	FD ý		

CP1252: Extended ASCII (winlatin1)



ISO 8859-2: Extended ASCII (latin2)



ISO 8859-3: Extended ASCII (latin3)

A0	A1 Ħ	A2 ˘	A3 £	A4 ¤		A6 Ĥ	A7 §	A8 ¨	A9 İ	AA Ş	AB Ğ	AC Ĵ	AD -		AF Ż
B0 °	B1 ħ	B2 ²	B3 ³	B4 ´	B5 µ	B6 ĥ	B7 ·	B8 ¸	B9 ı	BA ş	BB ğ	BC ĵ	BD ½		BF ż
C0 À	C1 Á	C2 Â		C4 Ä	C5 Ċ	C6 Ĉ	C7 Ç	C8 È	C9 É	CA Ê	CB Ë	CC Ì	CD Í	CE Î	CF Ï
	D1 Ñ	D2 Ò	D3 Ó	D4 Ô	D5 Ġ	D6 Ö	D7 ×	D8 Ĝ	D9 Ù	DA Ú	DB Û	DC Ü	DD Ŭ	DE Ŝ	DF ß
E0 à	E1 á	E2 â		E4 ä	E5 ċ	E6 ĉ	E7 ç	E8 è	E9 é	EA ê	EB ë	EC ì	ED í	EE î	EF ï
	F1 ñ	F2 ò	F3 ó	F4 ô	F5 ġ	F6 ö	F7 ÷	F8 ĝ	F9 ù	FA ú	FB û	FC ü	FD ŭ	FE ŝ	FF ˙

ISO 8859-4: Extended ASCII (latin4)

A0	A1	A2	A3	A4	A5	A6	A7	A8	A9	AA	AB	AC	AD	AE	AF
	Ą	ĸ	Ŗ	¤	Ĩ	Ļ	§	¨	Š	Ē	Ģ	Ŧ	-	Ž	¯
B0	B1	B2	B3	B4	B5	B6	B7	B8	B9	BA	BB	BC	BD	BE	BF
°	ą	˛	ŗ	´	ĩ	ļ	ˇ	¸	š	ē	ģ	ŧ	Ŋ	ž	ŋ
C0	C1	C2	C3	C4	C5	C6	C7	C8	C9	CA	CB	CC	CD	CE	CF
Ā	Á	Â	Ã	Ä	Å	Æ	Į	Č	É	Ę	Ë	Ė	Í	Î	Ī
D0	D1	D2	D3	D4	D5	D6	D7	D8	D9	DA	DB	DC	DD	DE	DF
Đ	Ņ	Ō	Ķ	Ô	Õ	Ö	×	Ø	Ų	Ú	Û	Ü	Ũ	Ū	ß
E0	E1	E2	E3	E4	E5	E6	E7	E8	E9	EA	EB	EC	ED	EE	EF
ā	á	â	ã	ä	å	æ	į	č	é	ę	ë	ė	í	î	ī
F0	F1	F2	F3	F4	F5	F6	F7	F8	F9	FA	FB	FC	FD	FE	FF
đ	ņ	ō	ķ	ô	õ	ö	÷	ø	ų	ú	û	ü	ũ	ū	˙

ISO 8859-5: Extended ASCII (Cyrillic)

A0	A1 Ё	A2 Ђ	A3 Ѓ	A4 Є	A5 Ѕ	A6 І	A7 Ї	A8 Ј	A9 Љ	AA Њ	AB Ћ	AC Ќ	AD	AE Ў	AF Џ
B0 А	B1 Б	B2 В	B3 Г	B4 Д	B5 Е	B6 Ж	B7 З	B8 И	B9 Й	BA К	BB Л	BC М	BD Н	BE О	BF П
C0 Р	C1 С	C2 Т	C3 У	C4 Ф	C5 Х	C6 Ц	C7 Ч	C8 Ш	C9 Щ	CA Ъ	CB Ы	CC Ь	CD Э	CE Ю	CF Я
D0 а	D1 б	D2 в	D3 г	D4 д	D5 е	D6 ж	D7 з	D8 и	D9 й	DA к	DB л	DC м	DD н	DE о	DF п
E0 р	E1 с	E2 т	E3 у	E4 ф	E5 х	E6 ц	E7 ч	E8 ш	E9 щ	EA ъ	EB ы	EC ь	ED э	EE ю	EF я
F0 №	F1 ё	F2 ђ	F3 ѓ	F4 є	F5 ѕ	F6 і	F7 ї	F8 ј	F9 љ	FA њ	FB ћ	FC ќ	FD §	FE ў	FF џ

ISO 8859-6: Extended ASCII (Arabic)

A0				A4 ¤								AC ،	AD ـ		
											BB ػ				BF ؿ
	C1 ء	C2 آ	C3 أ	C4 ؤ	C5 إ	C6 ئ	C7 ا	C8 ب	C9 ة	CA ت	CB ث	CC ج	CD ح	CE خ	CF د
D0 ذ	D1 ر	D2 ز	D3 س	D4 ش	D5 ص	D6 ض	D7 ط	D8 ظ	D9 ع	DA غ					
E0 ـ	E1 ف	E2 ق	E3 ك	E4 ل	E5 م	E6 ن	E7 ه	E8 و	E9 ى	EA ي	EB ً	EC ٌ	ED ٍ	EE َ	EF ُ
F0 ِ	F1 ّ	F2 ْ													

ISO 8859-7: Extended ASCII (Greek)

A0	A1 ‘	A2 ’	A3 £			A6 ¦	A7 §	A8 ¨	A9 ©		AB «	AC ¬	AD -		AF ―
B0 °	B1 ±	B2 ²	B3 ³	B4 ΄	B5 ΅	B6 Ά	B7 ·	B8 Έ	B9 Ή	BA Ί	BB »	BC Ό	BD ½	BE Ύ	BF Ώ
C0 ΐ	C1 Α	C2 Β	C3 Γ	C4 Δ	C5 Ε	C6 Ζ	C7 Η	C8 Θ	C9 Ι	CA Κ	CB Λ	CC Μ	CD Ν	CE Ξ	CF Ο
D0 Π	D1 Ρ		D3 Σ	D4 Τ	D5 Υ	D6 Φ	D7 Χ	D8 Ψ	D9 Ω	DA Ϊ	DB Ϋ	DC ά	DD έ	DE ή	DF ί
E0 ΰ	E1 α	E2 β	E3 γ	E4 δ	E5 ε	E6 ζ	E7 η	E8 θ	E9 ι	EA κ	EB λ	EC μ	ED ν	EE ξ	EF ο
F0 π	F1 ρ	F2 ς	F3 σ	F4 τ	F5 υ	F6 φ	F7 χ	F8 ψ	F9 ω	FA ϊ	FB ϋ	FC ό	FD ύ	FE ώ	

ISO 8859-8: Extended ASCII (Hebrew)

A0		A2 ¢	A3 £	A4 ¤	A5 ¥	A6 ¦	A7 §	A8 ¨	A9 ©	AA ×	AB «	AC ¬	AD	AE ®	AF ¯
B0 °	B1 ±	B2 ²	B3 ³	B4 ´	B5 µ	B6 ¶	B7 ·	B8 ¸	B9 ¹	BA ÷	BB »	BC ¼	BD ½	BE ¾	
															DF ‗
E0 א	E1 ב	E2 ג	E3 ד	E4 ה	E5 ו	E6 ז	E7 ח	E8 ט	E9 י	EA ך	EB כ	EC ל	ED ם	EE מ	EF ן
F0 נ	F1 ס	F2 ע	F3 ף	F4 פ	F5 ץ	F6 צ	F7 ק	F8 ר	F9 ש	FA ת					

ISO 8859-9: Extended ASCII (latin5)

A0	A1	A2	A3	A4	A5	A6	A7	A8	A9	AA	AB	AC	AD	AE	AF
B0	B1	B2	B3	B4	B5	B6	B7	B8	B9	BA	BB	BC	BD	BE	BF
C0	C1	C2	C3	C4	C5	C6	C7	C8	C9	CA	CB	CC	CD	CE	CF
D0	D1	D2	D3	D4	D5	D6	D7	D8	D9	DA	DB	DC	DD	DE	DF
E0	E1	E2	E3	E4	E5	E6	E7	E8	E9	EA	EB	EC	ED	EE	EF
F0	F1	F2	F3	F4	F5	F6	F7	F8	F9	FA	FB	FC	FD	FE	FF

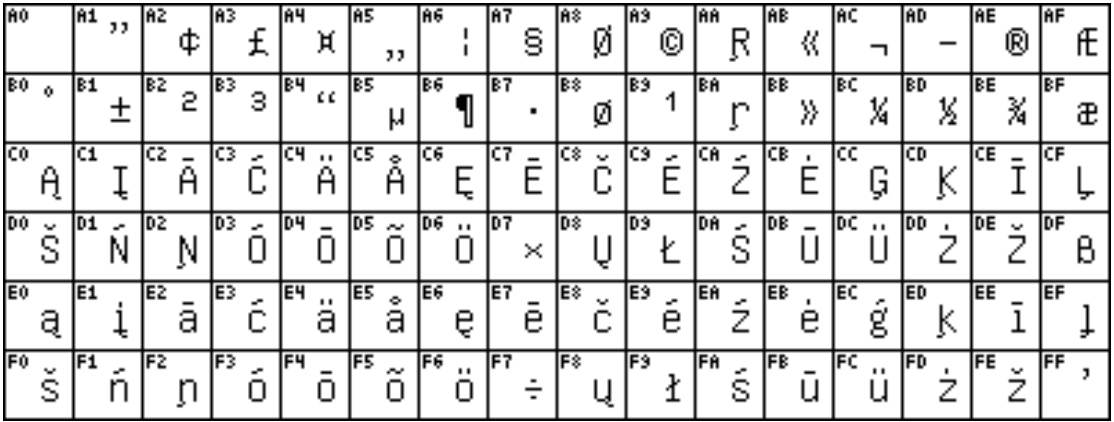
ISO 8859-10: Extended ASCII (latin6)

A0	A1	A2	A3	A4	A5	A6	A7	A8	A9	AA	AB	AC	AD	AE	AF
	Ą	Ē	Ģ	Ī	Ĩ	Ķ	§	Ļ	Đ	Š	Ŧ	Ž		Ū	Ŋ
B0	B1	B2	B3	B4	B5	B6	B7	B8	B9	BA	BB	BC	BD	BE	BF
°	ą	ē	ģ	ī	ĩ	ķ	·	ļ	đ	š	ŧ	ž	―	ū	ŋ
C0	C1	C2	C3	C4	C5	C6	C7	C8	C9	CA	CB	CC	CD	CE	CF
Ā	Á	Â	Ã	Ä	Å	Æ	Į	Č	É	Ę	Ë	Ė	Í	Î	Ï
D0	D1	D2	D3	D4	D5	D6	D7	D8	D9	DA	DB	DC	DD	DE	DF
Đ	Ņ	Ō	Ó	Ô	Õ	Ö	Ũ	Ø	Ų	Ú	Û	Ü	Ý	Þ	ß
E0	E1	E2	E3	E4	E5	E6	E7	E8	E9	EA	EB	EC	ED	EE	EF
ā	á	â	ã	ä	å	æ	į	č	é	ę	ë	ė	í	î	ï
F0	F1	F2	F3	F4	F5	F6	F7	F8	F9	FA	FB	FC	FD	FE	FF
ð	ņ	ō	ó	ô	õ	ö	ũ	ø	ų	ú	û	ü	ý	þ	ĸ

ISO 8859-11: Extended ASCII (Thai)

	A1	A2	A3	A4	A5	A6	A7	A8	A9	AA	AB	AC	AD	AE	AF
	ก	ข	ฃ	ค	ฅ	ฆ	ง	จ	ฉ	ช	ซ	ฌ	ญ	ฎ	ฏ
B0	B1	B2	B3	B4	B5	B6	B7	B8	B9	BA	BB	BC	BD	BE	BF
ฐ	ฑ	ฒ	ณ	ด	ต	ถ	ท	ธ	น	บ	ป	ผ	ฝ	พ	ฟ
C0	C1	C2	C3	C4	C5	C6	C7	C8	C9	CA	CB	CC	CD	CE	CF
ภ	ม	ย	ร	ฤ	ล	ฦ	ว	ศ	ษ	ส	ห	ฬ	อ	ฮ	ฯ
D0	D1	D2	D3	D4	D5	D6	D7	D8	D9	DA					DF
ะ	ั	า	ำ	ิ	ี	ึ	ื	ุ	ู	ฺ					฿
E0	E1	E2	E3	E4	E5	E6	E7	E8	E9	EA	EB	EC	ED	EE	EF
เ	แ	โ	ใ	ไ	ๅ	ๆ	็	่	้	๊	๋	์	ํ	๎	๏
F0	F1	F2	F3	F4	F5	F6	F7	F8	F9	FA	FB				
๐	๑	๒	๓	๔	๕	๖	๗	๘	๙	๚	๛				

ISO 8859-13: Extended ASCII (latin7)



ISO 8859-14: Extended ASCII (latin8)

A0	A1	A2	A3	A4	A5	A6	A7	A8	A9	AA	AB	AC	AD	AE	AF
B0	B1	B2	B3	B4	B5	B6	B7	B8	B9	BA	BB	BC	BD	BE	BF
C0	C1	C2	C3	C4	C5	C6	C7	C8	C9	CA	CB	CC	CD	CE	CF
D0	D1	D2	D3	D4	D5	D6	D7	D8	D9	DA	DB	DC	DD	DE	DF
E0	E1	E2	E3	E4	E5	E6	E7	E8	E9	EA	EB	EC	ED	EE	EF
F0	F1	F2	F3	F4	F5	F6	F7	F8	F9	FA	FB	FC	FD	FE	FF

ISO 8859-15: Extended ASCII (latin9)

A0	A1 ¡	A2 ¢	A3 £	A4 €	A5 ¥	A6 Š	A7 §	A8 š	A9 ©	AA ª	AB «	AC ¬	AD -	AE ®	AF ¯
B0 °	B1 ±	B2 ²	B3 ³	B4 Ž	B5 µ	B6 ¶	B7 ·	B8 ž	B9 ¹	BA º	BB »	BC Œ	BD œ	BE Ÿ	BF ¿
C0 À	C1 Á	C2 Â	C3 Ã	C4 Ä	C5 Å	C6 Æ	C7 Ç	C8 È	C9 É	CA Ê	CB Ë	CC Ì	CD Í	CE Î	CF Ï
D0 Ð	D1 Ñ	D2 Ò	D3 Ó	D4 Ô	D5 Õ	D6 Ö	D7 ×	D8 Ø	D9 Ù	DA Ú	DB Û	DC Ü	DD Ý	DE Þ	DF ß
E0 à	E1 á	E2 â	E3 ã	E4 ä	E5 å	E6 æ	E7 ç	E8 è	E9 é	EA ê	EB ë	EC ì	ED í	EE î	EF ï
F0 ð	F1 ñ	F2 ò	F3 ó	F4 ô	F5 õ	F6 ö	F7 ÷	F8 ø	F9 ù	FA ú	FB û	FC ü	FD ý	FE þ	FF ÿ

Historical PCs (around 1980)

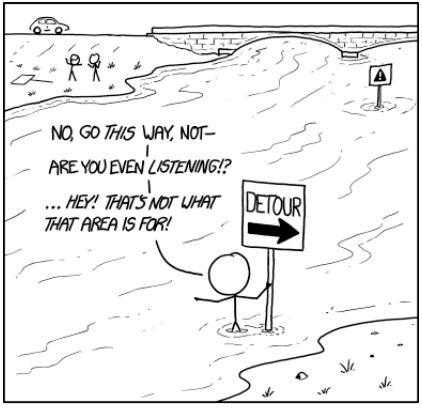
- Graphics card had a bitmap (say 9x14 bits) for each character
- Memory contained array with say 80x25 or 80x50 bytes for the text on the screen
- Graphics card would build video signal by indexing into the 256 bitmaps
- Firmware included initial bitmap used on system startup
- Different manufacturers used different mappings of numbers to characters

Major manufacturers create character encodings for major regions (1960-1991)

v · t · e	**Character encodings**	[hide]
Early telecommunications	ASCII • ISO/IEC 646 • ISO/IEC 6937 • T.61 • BCDIC • Baudot code • Morse code (Telegraph code • Wabun code) • Special telegraphy codes (Non-Latin • Chinese • Cyrillic) • Needle telegraph codes	
ISO/IEC 8859	-1 • -2 • -3 • -4 • -5 • -6 • -7 • -8 • -9 • -10 • -11 • -12 • -13 • -14 • -15 • -16	
Bibliographic use	ANSEL • ISO 5426 / 5426-2 / 5427 / 5428 / 6438 / 6861 / 6862 / 10585 / 10586 / 10754 / 11822 • MARC-8	
National standards	ArmSCII • BraSCII • CNS 11643 • ELOT 927 • GOST 10859 • GB 18030 • HKSCS • ISCII • JIS X 0201 • JIS X 0208 • JIS X 0212 • JIS X 0213 • KOI-7 • KPS 9566 • KS X 1001 • PASCII • SI 960 • TIS-620 • TSCII • VISCII • VSCII • YUSCII	
EUC	CN • JP • KR • TW	
ISO/IEC 2022	CN • JP • KR • CCCII	
MacOS code pages ("scripts")	Armenian • Arabic • Barents Cyrillic • Celtic • CentEuro • ChineseSimp / EUC-CN • ChineseTrad / Big5 • Croatian • Cyrillic • Devanagari • Dingbats • Esperanto • Farsi (Persian) • Gaelic • Georgian • Greek • Gujarati • Gurmukhi • Hebrew • Iceland • Inuit • Japanese / ShiftJIS • Keyboard • Korean / EUC-KR • Latin-1 • Ogham • Roman • Romanian • Sámi • Symbol • Thai / TIS-620 • Turkish • Turkic Latin • Turkic Cyrillic • Ukrainian	
DOS code pages	100 • 111 • 112 • 113 • 151 • 152 • 161 • 162 • 163 • 164 • 165 • 166 • 210 • 220 • 301 • 437 • 449 • 489 • 620 • 667 • 668 • 707 • 708 • 709 • 710 • 711 • 714 • 715 • 720 • 721 • 737 • 768 • 770 • 771 • 772 • 773 • 774 • 775 • 776 • 777 • 778 • 790 • 850 • 851 • 852 • 853 • 854 • 855/872 • 856 • 857 • 858 • 859 • 860 • 861 • 862 • 863 • 864/17248 • 865 • 866/808 • 867 • 868 • 869 • 874/1161/1162 • 876 • 877 • 878 • 881 • 882 • 883 • 884 • 885 • 891 • 895 • 896 • 897 • 898 • 899 • 900 • 903 • 904 • 906 • 907 • 909 • 910 • 911 • 926 • 927 • 928 • 929 • 932 • 934 • 936 • 938 • 941 • 942 • 943 • 944 • 946 • 947 • 948 • 949 • 950/1370 • 951 • 966 • 991 • 1034 • 1039 • 1040 • 1041 • 1042 • 1043 • 1044 • 1046 • 1086 • 1088 • 1092 • 1093 • 1098 • 1108 • 1109 • 1114 • 1115 • 1116 • 1117 • 1118 • 1119 • 1125/848 • 1126 • 1127 • 1131/849 • 1139 • 1167 • 1168 • 1300 • 1351 • 1361 • 1362 • 1363 • 1372 • 1373 • 1374 • 1375 • 1380 • 1381 • 1385 • 1386 • 1391 • 1392 • 1393 • 1394 • CWI-2 • Iran System • Kamenický • KOI8 • Mazovia • MIK	
IBM AIX code pages	367 • 371 • 806 • 813 • 819 • 895 • 896 • 912 • 913 • 914 • 915 • 916 • 919 • 920 • 921/901 • 922/902 • 923 • 952 • 953 • 954 • 955 • 956 • 957 • 958 • 959 • 960 • 961 • 963 • 964 • 965 • 970 • 971 • 1004 • 1006 • 1008 • 1009 • 1010 • 1011 • 1012 • 1013 • 1014 • 1015 • 1016 • 1017 • 1018 • 1019 • 1029 • 1036 • 1089 • 1111 • 1124 • 1129/1163 • 1133 • 1350 • 1382 • 1383	
IBM Apple Macintosh emulations	1275 • 1280 • 1281 • 1282 • 1283 • 1284 • 1285 • 1286	
IBM Adobe emulations	1038 • 1276 • 1277	
IBM DEC emulations	1020 • 1021 • 1023 • 1090 • 1100 • 1101 • 1102 • 1103 • 1104 • 1105 • 1106 • 1107 • 1287 • 1288	
IBM HP emulations	1050 • 1051 • 1052 • 1053 • 1054 • 1055 • 1056 • 1057 • 1058	
Windows code pages	CER-GS • 874/1162 (TIS-620) • 932/943 (Shift JIS) • 936/1386 (GBK) • 950/1370 (Big5) • 949/1363 (EUC-KR) • 1169 • 1174 • Extended Latin-8 • 1200 (UTF-16LE) • 1201 (UTF-16BE) • 1250 • 1251 • 1252 • 1253 • 1254 • 1255 • 1256 • 1257 • 1258 • 1259 • 1261 • 1270 • 54936 (GB18030)	
EBCDIC code pages	1 • 2 • 3 • 4 • 5 • 6 • 7 • 8 • 9 • 10 • 11 • 12 • 13 • 14 • 15 • 16 • 17 • 18 • 19 • 20 • 21 • 22 • 23 • 24 • 25 • 26 • 27 • 28 • 29 • 30 • 31 • 32 • 33 • 34 • 35 • 36 • 37/1140 • 37-2 • 38 • 39 • 40 • 251 • 252 • 254 • 256 • 257 • 258 • 259 • 260 • 264 • 273/1141 • 274 • 275 • 276 • 277/1142 • 278/1143 • 279 • 280/1144 • 281 • 282 • 283 • 284/1145 • 285/1146 • 286 • 287 • 288 • 289 • 290 • 293 • 297/1147 • 298 • 300 • 310 • 320 • 321 • 322 • 330 • 351 • 352 • 353 • 355 • 357 • 358 • 359 • 360 • 361 • 363 • 382 • 383 • 384 • 385 • 386 • 387 • 388 • 389 • 390 • 391 • 392 • 393 • 394 • 395 • 410 • 420/16804 • 421 • 423 • 424/8616/12712 • 425 • 435 • 500/1148 • 803 • 829 • 833 • 834 • 835 • 836 • 837 • 838/838 • 839 • 870/1110/1153 • 871/1149 • 875/4971/9067 • 880 • 881 • 882 • 883 • 884 • 885 • 886 • 887 • 888 • 889 • 890 • 892 • 893 • 905 • 918 • 924 • 930/1390 • 931 • 933/1364 • 935/1388 • 937/1371 • 939/1399 • 1001 • 1002 • 1003 • 1005 • 1007 • 1024 • 1025/1154 • 1026/1155 • 1027 • 1028 • 1030 • 1031 • 1032 • 1033 • 1037 • 1047 • 1068 • 1069 • 1070 • 1071 • 1073 • 1074 • 1075 • 1076 • 1077 • 1078 • 1079 • 1080 • 1081 • 1082 • 1083 • 1084 • 1085 • 1087 • 1091 • 1097 • 1112/1156 • 1113 • 1122/1157 • 1123/1158 • 1130/1164 • 1132 • 1136 • 1137 • 1150 • 1151 • 1152 • 1159 • 1165 • 1166 • 1278 • 1279 • 1303 • 1364 • 1376 • 1377 • JEF • KEIS	
Platform specific	Acorn • Adobe Standard • Adobe Latin 1 • Apple II • ATASCII • Atari ST • BICS • Casio calculators • CDC • CPC • DEC Radix-50 • DEC MCS/NRCS • DG International • ELWRO-Junior • FIELDATA • GEM • GEOS • GSM 03.38 • HP Roman Extension • HP Roman-8 • HP Roman-9 • HP FOCAL • HP RPL • LICS • LMBCS • Mattel Aquarius • MSX • NEC APC • NeXT • PCW • PETSCII • Sharp calculators • TI calculators • TRS-80 • Ventura International • Ventura Symbol • WISCII • XCCS • ZX80 • ZX81 • ZX Spectrum	
Unicode / ISO/IEC 10646	UTF-1 • UTF-7 • UTF-8 • UTF-16 (UTF-16LE/UTF-16BE) / UCS-2 • UTF-32 (UTF-32LE/UTF-32BE) / UCS-4 • UTF-EBCDIC • GB 18030 • BOCU-1 • CESU-8 • SCSU	
Miscellaneous code pages	ABICOMP • APL • ARIB STD-B24 • Cork • HZ • INIS • INIS-8 • ISO-IR-111 • ISO-IR-182 • ISO-IR-197 • ISO-IR-200 • ISO-IR-201 • Johab • LGR • LY1 • OML • OMS • OMX • OT1 • OT2 • OT3 • OT4 • T2A • T2B • T2C • T2D • T3 • T4 • T5 • TS1 • TS3 • U • X2 • SEASCII • TACE16 • TRON • UTF-5 • UTF-6 • WTF-8	
Related topics	Code page • Control character (C0 C1) • CCSID • Character encodings in HTML • Charset detection • Han unification • Hardware • ISO 6429/IEC 6429/ANSI X3.64 • Mojibake	
	Character sets	

Unicode

Mission impossible



WATCHING THE UNICODE PEOPLE TRY TO GOVERN THE INFINITE CHAOS OF HUMAN LANGUAGE WITH CONSISTENT TECHNICAL STANDARDS IS LIKE WATCHING HIGHWAY ENGINEERS TRY TO STEER A RIVER USING TRAFFIC SIGNS.

1991: We urgently need a standard for characters…

- ISO 10646 project
- US consortium: Unicode project

Fortunately:

Unicode 1.1 = ISO 10646-1:1993
Unicode 3.0 = ISO 10646-1:2000
Unicode 7.0 = ISO 10646:2014
etc.

… and all are backwards-compatible since Unicode 2.0.

Unicode uses a 5 layer architecture

1. Abstract Character Repertoire
2. Coded Character Set
3. Character Encoding Form
4. Character Encoding Scheme
5. Transfer Encoding Syntax

Abstract character repertoire (ACR)

An ACR specifies:

- set of unique characters
- each character has a general name (“at sign”)
- each character has a graphical representation (“@”)

It leaves open:

- representation of characters
- order of characters

Coded character set (CCS)

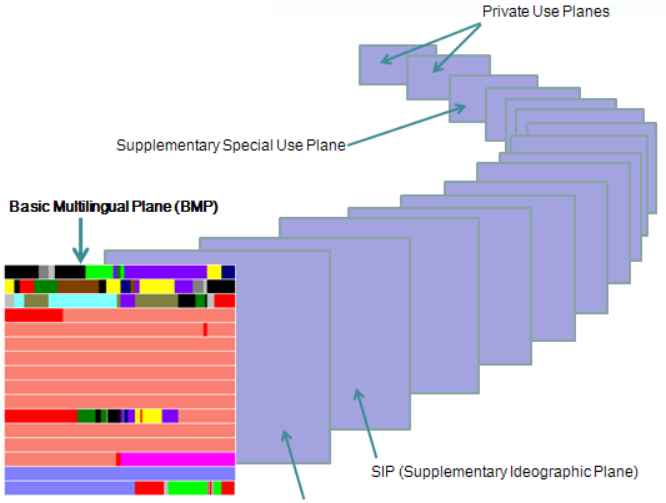
- 1:1 mapping of ACR to positive numbers
- each coded character has a numeric code
- each coded character has a standardized name (“COMMERCIAL AT”)
- each coded character has a code position

Note that a CCS can have holes with positions left unspecified.

`http://www.unicode.org/charts/`

Basic Multilingual Plane (BMP)

First 65536 code points of Unicode are called the BMP:



Character Encoding Form (CEF)

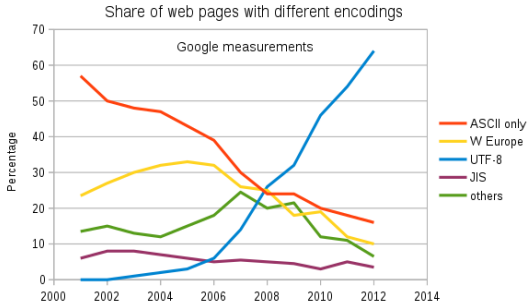
- Specifies mapping of code numbers to code units
- Code units are bit sequences of **fixed length** (usually 8, 16, 24, 32, etc.)

Character Encoding Scheme (CES)

- Mapping from code units to serialized byte sequences
- Byte order (big endian, little endian) matters
- Examples: UTF-16BE vs UTF-16LE[a]

UTF-8 is the most widely used CES.

[a] UTF = UCS transformation format, UCS = Universal Coded-Character Set



Example

Name	Latin A	Hebrew alef	Han AN
Code point	U+0041	U+05D0	U+597D
UTF-8	41H	D7H 90H	E5H A5H BDH
UTF-16BE	00 41H	5H D0H	59H 7DH
UTF-32BE	00 00 00 41H	00 00 5H D0H	00 00 59H 7DH

Transfer Encoding Syntax (TES)

Invertible conversion of byte sequences to:

- eliminate certain undesirable byte sequences that might confuse transfer protocols
- reduce bandwidth consumption via compression (gzip, deflate)

Example

1. ACR specifies collection of characters, i.e. a , !, ä and ‰.
2. CCS specifies numeric codes, i.e. ISO 10646 uses 97, 33, 228 and 8240 (0x2030) for the characters above.
3. A CEF specifies that the above codes are represented using two bytes (UCS-2) or four bytes (UCS-4).
4. A CES may specify how the two bytes are encoded, i.e in big-endian, i.e. for UTF-16BE: (0;97), (0;33), (0;228), (32;48) = (0x20;0x30).
5. A TES may for example apply `gzip` compression to the above sequence.

UTF-8

UTF-8: Motivation

- Most transmitted characters still from 7-bit ASCII
- Common characters in text are in BMP
- Using 32 bits per character would be inefficient!
- C programming language does not deal well with 0 in strings!

UTF-8: Rules

- 7-bit ASCII characters use one byte (0xxxxxxx)
- Up to 2^{11} code points use two bytes (110xxxxx.10xxxxxx)
- Up to 2^{16} code points use three bytes (1110xxxx.10xxxxxx.10xxxxxx)
- Up to 2^{21} code points use four bytes (11110xxx.10xxxxxx. 10xxxxxx.10xxxxxx)
- Etc.[1]

Note: Starting with version 3.1, Unicode requires the shortest possible representation!

oxFE and oxFF cannot appear in UTF-8.
oxoo only appears for o.

[1] While this encoding supports up to six-byte values in theory, only up to four bytes are allowed in Unicode.

UTF-8: Properties

- Can encode all 2^{21} UCS characters
- Characters may take up to 4 octets, `strlen()` will not work!
- Strings sorted using UCS-4BE will remain sorted
- Applications **may** reject encodings that use more space than required