



Berner Fachhochschule  
Haute école spécialisée bernoise  
Bern University of Applied Sciences

# CS Basics

## 1) Character Sets and Encodings

*E. Benoist & C. Grothoff*  
Fall Term 2018-19



# Agenda

# 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

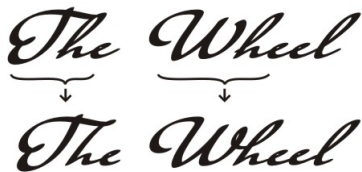


*"A grapheme (computing: character) is the smallest unit of a writing system of any given language."*

*–<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 (1830)

	Five-needle	Two-needle	One-needle
A	/     \	//	// (7)
B	/     \	///	/// (8)
C		✓   (1)	//// (1)
D	/     \	✓   (2)	✓ \ (9)
E	/   \	/   (3)	✓ \ (0)
F	/   \	//	/// \
G	/   \	///	✓ \ (+)
H	/ \	\ (4)	✓ \
I	/ \	//	✓ \
J		(subst. G)	
K	/ \	///	✓ \ (Wait)
L	/ \	\ (5)	✓ \ (Express)
M	\ \	✓ (6)	/ (1)
N	\ /	/ (7)	// (2)
O	\ /	//	/// (3)
P	\ /	///	//// (4)
Q		(subst. K)	
R	\ \ /	\ \ (8)	✓ (5)
S	\   /	// \	✓ \ (6)
T	\   /	/// \	/// \
U	\ \     /	✓ \ (9)	✓ \
V		✓ \ (0)	✓ \

# Extended Binary Coded Decimal Interchange Code (1963)

2nd hex digit

EBCDIC character codes

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

Berner Fachhochschule Bern University of Applied Sciences

# US ASCII (1963)

## ASCII TABLE

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

# Control codes

Character sets often contain control codes, code positions that are not mapped to a visible character but for 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





# ISO 8859-1: Extended ASCII (latin1)

	A1	A2	A3		A5		A7	A8	A9	AA	AB					
	í	í	£		¥		§	¤	©	≡	«					
B0	°	B1	B2	B3	B5	B6	B7		B9	1	»	¼	½		¿	
	±	²	³		µ	¶	·		¹	º	»	¼	½		¿	
C0	À	C1	C2	C3	C4	C5	C6	C7	C8	C9	CA	CB	CC	CD	CE	CF
	À	Á	Â	Ã	Ä	Å	Æ	Ç	È	É	Ê	Ë	Ì	Í	Î	Ï
	Ñ	Ò	Ó	Ô	Õ	Ö	Ø	Ù	Ú	Û	Ü	Ý				Þ
E0	à	E1	E2	E3	E4	E5	E6	E7	E8	E9	EA	EB	EC	ED	EE	EF
	à	á	â	ã	ä	å	æ	ç	è	é	ê	ë	ì	í	î	ï
	ñ	ò	ó	ô	õ	ö	ø	ù	ú	û	ü	ý				

# CP1252: Extended ASCII (winlatin1)

80	€		82	,	83	f	84	,,	85	...	86	†	87	‡	88	~	89	%	8A	š	8B	<	8C	œ		8E	ž				
	91	‘	92	’	93	“	94	”	95	•	96	-	97	-	98	~	99	™	9A	š	9B	>	9C	œ		9E	ž	9F	ÿ		
A0	A1	ı	A2	ϕ	A3	£	A4	¤	A5	¥	A6	ı	A7	§	A8	..	A9	©	AA	≡	AB	«	AC	¬	AD	-	AE	®	AF	-	
B0	°	B1	±	B2	²	B3	³	B4	ˆ	B5	µ	B6	¶	B7	·	B8	,	B9	¹	BA	º	BB	»	BC	¼	BD	½	BE	¾	BF	¿
C0	À	Á	Â	Ã	Ä	Å	Æ	Ç	È	É	Ê	Ë	Ì	Í	Î	Ï															
D0	Ð	Ñ	Ò	Ó	Ô	Õ	Ö	×	Ø	Ù	Ú	Û	Ü	Ý	Þ	ß															
E0	à	á	â	ã	ä	å	æ	ç	è	é	ê	ë	ì	í	î	ï															
F0	ð	ñ	ò	ó	ô	õ	ö	÷	ø	ù	ú	û	ü	ý	þ	ÿ															

# ISO 8859-2: Extended ASCII (latin2)

A0	A1 Ā	A2 ˘	A3 Ł	A4 Ȩ	A5 Ĺ	A6 Š	A7 Š	A8 ˙	A9 Š	AA Š	AB Ț	AC Ž	AD -	AE Ž	AF Ž
B0 ˚	B1 ā	B2 ˘	B3 ł	B4 Ȩ	B5 Ĺ	B6 š	B7 ˘	B8 ˙	B9 š	BA š	BB ț	BC ž	BD ˘	BE ž	BF ž
C0 Ā	C1 Ā	C2 Ā	C3 Ā	C4 Ā	C5 Ĺ	C6 Ć	C7 Ć	C8 Ć	C9 Ę	CA Ę	CB Ę	CC Ę	CD Ī	CE Ī	CF Ď
D0 Đ	D1 Ñ	D2 Ñ	D3 Ñ	D4 Ñ	D5 Ñ	D6 Ñ	D7 ×	D8 Ř	D9 Ů	DA Ů	DB Ů	DC Ů	DD Ý	DE J	DF B
E0 ř	E1 ā	E2 â	E3 ä	E4 ä	E5 Ĺ	E6 Ć	E7 Ć	E8 Ć	E9 ě	EA ě	EB ě	EC ě	ED í	EE î	EF ě
F0 đ	F1 ñ	F2 ñ	F3 ñ	F4 ñ	F5 ñ	F6 ñ	F7 ÷	F8 ř	F9 Ů	FA Ů	FB Ů	FC Ů	FD ý	FE †	FF ˙

# ISO 8859-3: Extended ASCII (latin3)

A0	A1	A2	A3	A4		A6	A7	A8	A9	AA	AB	AC	AD		AF	
	Ħ	ı	£	₣		Ĥ	Š	..	İ	Ş	Ğ	Ĵ	-		Ž	
B0	◊	ħ	2	3	ı	ĥ	•	˙	ı	š	ğ	ĵ	¼		ž	
C0	Ã	Ä	Å		Ä	Č	Ĉ	Ç	È	É	Ê	Ë	Ī	Ĭ	Ī	İ
	D1	Ñ	Õ	Ö	Ô	Ġ	Ö	×	Ĝ	Û	Ū	Ū	Ü	Ŭ	Ŝ	ß
E0	ã	ä	å		ä	č	ĉ	ç	è	é	ê	ë	ĩ	î	ï	ï
	F1	ñ	õ	ö	ô	ğ	ö	÷	ğ	û	ü	ü	ü	ŭ	ŝ	·

# ISO 8859-4: Extended ASCII (latin4)

A0	A1	A2	A3	A4	A5	A6	A7	A8	A9	AA	AB	AC	AD	AE	AF
	À	Ā	Ŕ	Ȩ	İ̇	Ł	Ś	..	Š	Ē	Ġ	Ʀ	-	Ž	-
B0	á	ă	ŗ	ȩ	ĩ̇	ł	ś	..	š	ē	ġ	ƣ	ŋ	ž	ŋ
C0	Ā	Ă	Â	Ã	Ä	Å	Æ	İ	Č	É	Ě	Ë	Ī	Î	Ĭ
D0	Đ	Ń	Ō	Ķ	Ô	Õ	Ö	×	Ø	Ū	Ū	Û	Ü	Ŭ	Ů
E0	ā	ă	â	ã	ä	å	æ	ı̇	č	é	ě	ë	î	ï	ĭ
F0	đ	ñ	õ	ķ	ô	õ	ö	÷	ø	ų	ŭ	ű	ü	ŭ	·

# ISO 8859-5: Extended ASCII (Cyrillic)

A0	A1	A2	A3	A4	A5	A6	A7	A8	A9	AA	AB	AC	AD	AE	AF
	Ё	Ђ	Ѓ	Є	Ѕ	І	Ї	Ј	Љ	Њ	Ћ	Ќ	–	Ў	Ў
B0	В	В	Г	Д	Е	Ж	З	И	Й	К	Л	М	Н	О	П
C0	Р	С	Т	У	Ф	Х	Ц	Ч	Ш	Щ	Ъ	Ы	Ь	Э	Ю
D0	а	б	в	г	д	е	ж	з	и	й	к	л	м	н	о
E0	р	с	т	у	ф	х	ц	ч	ш	щ	ъ	ы	ь	э	ю
F0	ё	ђ	ѓ	є	ѕ	і	ї	ј	љ	њ	ћ	ќ	ѕ	ў	џ

# ISO 8859-6: Extended ASCII (Arabic)

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

# ISO 8859-7: Extended ASCII (Greek)

A0	A1	A2	A3			A6	A7	A8	A9		AE	AC	AD		AF
	€	‚	£			ı	š	¨	©		«	¬	-		-
B0	B1	B2	B3	B4	B5	B6	B7	B8	B9	BA	BB	BC	BD	BE	BF
◊	±	²	³	ı	ˆ	À	•	É	Ĥ	Ì	»	Ò	¼	Ý	Ω
C0	C1	C2	C3	C4	C5	C6	C7	C8	C9	CA	CB	CC	CD	CE	CF
ı̇	À	Β	Γ	Δ	Ε	Ζ	Η	Θ	Ι	Κ	Λ	Μ	Ν	Ξ	Ο
D0	D1		D3	D4	D5	D6	D7	D8	D9	DA	DB	DC	DD	DE	DF
Π	Ρ		Σ	Τ	Υ	Φ	Χ	Ψ	Ω	İ	ÿ	ά	έ	ή	ι
E0	E1	E2	E3	E4	E5	E6	E7	E8	E9	EA	EB	EC	ED	EE	EF
ı̇	α	β	γ	δ	ε	ζ	η	θ	ι	κ	λ	μ	ν	ξ	ο
F0	F1	F2	F3	F4	F5	F6	F7	F8	F9	FA	FB	FC	FD	FE	
π	ρ	ς	σ	τ	υ	φ	χ	ψ	ω	ı̇	ÿ	ó	ú	ώ	



# ISO 8859-8: Extended ASCII (Hebrew)

A0		A2	Φ	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	º	BB	»	BC	¼	BD	½	BE	¾		
																												DF	=		
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	Ĵ	F1	Đ	F2	ǰ	F3	ǧ	F4	ǥ	F5	Ƴ	F6	Ʒ	F7	Ɔ	F8	Ɣ	F9	Ʊ	FA	Ƭ										

# ISO 8859-9: Extended ASCII (latin5)

A0	A1	A2	A3	A4	A5	A6	A7	A8	A9	AA	AB	AC	AD	AE	AF
	ı	ϕ	£	¥	¥	ı	š	..	©	≡	«	¬	-	®	-
B0	°	±	²	³	ˆ	µ	¶	·	¸	¹	º	»	¼	½	¾
C0	À	Á	Â	Ã	Ä	Å	Æ	Ç	È	É	Ê	Ë	Ì	Í	Î
D0	Ĝ	Ñ	Ò	Ó	Ô	Õ	Ö	×	Ø	Ù	Ú	Û	Ü	İ	Ş
E0	à	á	â	ã	ä	å	æ	ç	è	é	ê	ë	ì	í	î
F0	ğ	ñ	ò	ó	ô	õ	ö	÷	ø	ù	ú	û	ü	ı	ş

# ISO 8859-10: Extended ASCII (latin6)

A0	A1	A2	A3	A4	A5	A6	A7	A8	A9	AA	AB	AC	AD	AE	AF
	À	Ā	Ġ	Ī	Ĩ	Ɔ	Š	Ł	Đ	Š̂	Ʀ	Ž̂	-	Ū	Ŋ
B0	á	ē	ġ	ī	ĩ	Ɔ	•	ł	đ	š̂	Ʀ	ž̂	-	ū	ŋ
C0	Ā	Ā̂	Ā̂	Ā̂	Ā̂	Ā̂	Æ	ı	č	é	ę	ë	è	í	î
D0	ð	Ń	ō	õ	ô	õ	ö	ũ	ø	ų	ú	û	ü	ý	þ
E0	ā	ā̂	ā̂	ā̂	ā̂	æ	ı	č	é	ę	ë	è	í	î	ï
F0	ð	Ń	ō	õ	ô	õ	ö	ũ	ø	ų	ú	û	ü	ý	þ

# 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	ฉ	ช	ฌ	ฉ	ซ	ฌ	ฌ	ฌ	ฌ	ฌ
B0	B1	B2	B3	B4	B5	B6	B7	B8	B9	BA	BB	BC	BD	BE	BF
๒	ก	ข	ฃ	ค	ฅ	ฉ	ช	ฌ	ฉ	ซ	ฌ	ฌ	ฌ	ฌ	ฌ
C0	C1	C2	C3	C4	C5	C6	C7	C8	C9	CA	CB	CC	CD	CE	CF
ก	ข	ฃ	ค	ฅ	ฉ	ช	ฌ	ฉ	ซ	ฌ	ฌ	ฌ	ฌ	ฌ	ฌ
D0	D1	D2	D3	D4	D5	D6	D7	D8	D9	DA					DF
๕	๖	๗	๘	๙	๑	๒	๓	๔	๕	๖					฿
E0	E1	E2	E3	E4	E5	E6	E7	E8	E9	EA	EB	EC	ED	EE	EF
๑	๒	๓	๔	๕	๖	๗	๘	๙	๑	๒	๓	๔	๕	๖	๗
F0	F1	F2	F3	F4	F5	F6	F7	F8	F9	FA	FB				
๐	๑	๒	๓	๔	๕	๖	๗	๘	๙	๐	๑				

# ISO 8859-13: Extended ASCII (latin7)

A0	A1	A2	A3	A4	A5	A6	A7	A8	A9	AA	AB	AC	AD	AE	AF	
	„	φ	£	℥	„	!	§	∅	©	℞	«	¬	-	®	æ	
B0	°	±	²	³	“	μ	¶	•	∅	¹	ℓ	»	¼	½	¾	æ
C0	À	Ā	Ā	Ä	Å	Ē	Ē	Č	É	Ž	È	Ĝ	Ķ	Ī	Ĵ	
D0	Š	Ń	Ń	Ō	Ō	Ö	×	Ū	Ł	Ś	Ū	Ü	Ż	Ž	ß	
E0	ą	ı	ā	ć	ä	å	ę	ē	č	é	ž	è	ğ	ķ	ī	ĵ
F0	š	ń	ń	ō	ō	ö	÷	ū	ł	ś	ū	ü	ż	ž	,	

# ISO 8859-14: Extended ASCII (latin8)

A0	A1	A2	A3	A4	A5	A6	A7	A8	A9	AA	AB	AC	AD	AE	AF	
	À	á	â	Ë	Ċ	Ċ	Ď	Š	Ŧ	©	Ŧ	đ	ÿ	-	®	ÿ
B0	ƒ	ġ	ġ	Ĥ	Ĥ	¶	Ŗ	ŵ	Ŗ	ŵ	š	ÿ	ŵ	ŵ	š	
C0	À	Ā	Ā	Ä	Ä	Æ	Ç	È	É	Ê	Ë	Ī	Ī	Ī	Ī	
D0	Ŧ	Ŧ	Ŧ	Ŧ	Ŧ	Ŧ	Ŧ	Ŧ	Ŧ	Ŧ	Ŧ	Ŧ	Ŧ	Ŧ	Ŧ	Ŧ
E0	à	ā	ā	ä	ä	æ	ç	è	é	ê	ë	ĭ	ĭ	ĭ	ĭ	
F0	ŵ	Ŧ	Ŧ	Ŧ	Ŧ	Ŧ	Ŧ	Ŧ	Ŧ	Ŧ	Ŧ	Ŧ	Ŧ	Ŧ	Ŧ	Ŧ

# ISO 8859-15: Extended ASCII (latin9)

A0	A1	A2	A3	A4	A5	A6	A7	A8	A9	AA	AB	AC	AD	AE	AF
	ı	ϕ	£	€	¥	Š	š	Š	©	≡	«	¬	-	®	-
B0	°	±	²	³	Ž	μ	¶	·	ž	ı	ó	»	œ	œ	ÿ
C0	À	Á	Â	Ã	Ä	Å	Æ	Ç	È	É	Ê	Ë	Ì	Í	Î
D0	Ð	Ñ	Ò	Ó	Ô	Õ	Ö	×	Ø	Ù	Ú	Û	Ü	Ý	Þ
E0	à	á	â	ã	ä	å	æ	ç	è	é	ê	ë	ì	í	î
F0	ð	ñ	ò	ó	ô	õ	ö	÷	ø	ù	ú	û	ü	ý	þ

# 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]
<b>Early telecommunications</b>	ASCII · ISO/IEC 646 · ISO/IEC 6937 · T.61 · BCDC · Baudot code · Morse code (Telegraph code · Wabun code) · Special telegraphy codes (Non-Latin · Chinese · Cyrillic) · Needle telegraph codes	
<b>ISO/IEC 8859</b>	-1 · -2 · -3 · -4 · -5 · -6 · -7 · -8 · -9 · -10 · -11 · -12 · -13 · -14 · -15 · -16	
<b>Bibliographic use</b>	ANSEL · ISO 5426 / 5426-2 / 5427 / 5428 / 6438 / 6961 / 6962 / 10585 / 10586 / 10754 / 11822 · MARC-8	
<b>National standards</b>	AmSCII · 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 · KRSCII · SI 960 · TIS-620 · TSCII · VISCI · VSCII · YLUSCI	
<b>EUC</b>	CN · JP · KR · TW	
<b>ISO/IEC 2022</b>	CN · JP · KR · CCCII	
<b>MacOS code pages ("scripts")</b>	Armenian · Arabic · Barents Cyrillic · Celtic · CentEuro · ChineseSimp / EUC-CN · ChineseTrad / Big5 · Croatian · Cyrillic · Devanagari · Dingbats · Esperanto · Farsi (Persian) · Gaelic · Georgian · Greek · Gujarati · Gurmukhi · Hebrew · Icelandic · Inuit · Japanese / ShiftJIS · Keyboard · Korean / EUC-KR · Latin-1 · Ogham · Roman · Romanian · Sâmi · Symbol · Thai / TIS-620 · Turkish · Turkic Latin · Turkic Cyrillic · Ukrainian	
<b>DOS code pages</b>	100 · 111 · 112 · 113 · 115 · 125 · 162 · 163 · 164 · 165 · 166 · 210 · 220 · 301 · 437 · 449 · 489 · 620 · 667 · 668 · 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 · 864/1248 · 865 · 866/808 · 867 · 868 · 869 · 874/1611/162 · 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 · Kamenicky · KOI8 · Mazovia · MIK	
<b>IBM AIX code pages</b>	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 · 1362 · 1363	
<b>IBM Apple Macintosh emulations</b>	1275 · 1280 · 1281 · 1282 · 1283 · 1284 · 1285 · 1286	
<b>IBM Adobe emulations</b>	1038 · 1276 · 1277	
<b>IBM DEC emulations</b>	1020 · 1021 · 1023 · 1090 · 1100 · 1101 · 1102 · 1103 · 1104 · 1105 · 1106 · 1107 · 1187 · 1288	
<b>IBM HP emulations</b>	1050 · 1051 · 1052 · 1053 · 1054 · 1055 · 1056 · 1057 · 1058	
<b>Windows code pages</b>	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 · 1272 · 54936 (GB18030)	
<b>EBCDIC code pages</b>	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 · 273/141 · 274 · 275 · 276 · 277/142 · 278/143 · 279 · 280/144 · 281 · 282 · 283 · 284/145 · 285/146 · 286 · 287 · 288 · 289 · 290 · 285 · 297/147 · 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 · 424/816/12712 · 425 · 435 · 500/1148 · 803 · 829 · 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 · 895 · 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 · 1112/1157 · 1123/1158 · 1130/1164 · 1132 · 1136 · 1137 · 1150 · 1151 · 1152 · 1159 · 1165 · 1166 · 1178 · 1279 · 1303 · 1364 · 1376 · 1377 · JEF · KEIS	
<b>Platform specific</b>	Accorn · Adobe Standard · Adobe Latin 1 · Apple II · ATASCII · Atari ST · BCIS · 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 · UCS · UMBCS · Mattel Aquarius · MSX · NEC APC · NeXT · PCW · PETSCII · Sharp calculators · TI calculators · TIS-80 · Ventura International · Ventura Symbol · VISCI · XCS · ZX80 · ZX81 · ZX Spectrum	
<b>Unicode / ISO/IEC 10646</b>	UTF-1 · UTF-7 · UTF-8 · UTF-16 (UTF-16LE/UTF-16BE) / UCS-2 · UTF-32 (UTF-32LE/UTF-32BE) / UCS-4 · UTF-8BICD · GB 18030 · BOCU-1 · CESU-8 · SCSU	
<b>Miscellaneous code pages</b>	ARIBCP · APL · ARIB STD B24 · Carik · HZ · IMS · IMS-8 · ISO-IR-111 · ISO-IR-182 · ISO-IR-197 · ISO-IR-200 · ISO-IR-201 · jahab · LGR · L71 · OML · OMS · OMX · OT1 · OT2 · OT3 · OT4 · T2A · T2B · T2C · T2D · T3 · T4 · T5 · T51 · T53 · U · X2 · XEASCII · TACE1A · TRON · UTF-5 · UTF-7 · WTF-8	
<b>Related topics</b>	Code page · Control character (C0 C1) · CCSID · Character encodings in HTML · Charset detection · Han unification · Hardware · ISO 6429/REC 6429/JANS X3.64 · Mojibake	

[Character sets](#)

# Unicode

# 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

# 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

... 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)

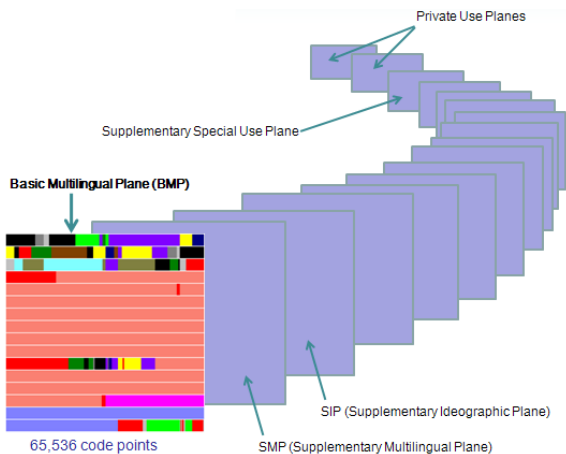
- ▶ 1:1 mapping of ACR to positive numbers
- ▶ each coded character has a numeric code
- ▶ each coded character has a standardized name (“COMMERCIAL AT”)
- ▶ each coded character has a code position

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:



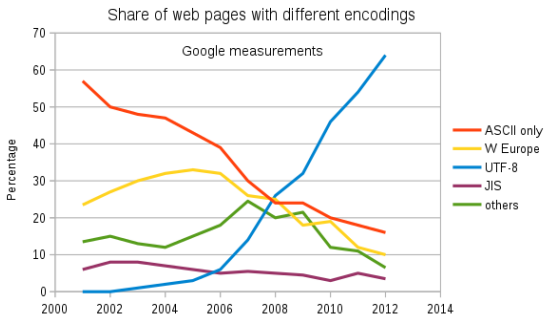
# Character Encoding Form (CEF)

- ▶ 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<sup>1</sup>

UTF-8 is the most widely used CES:



<sup>1</sup>UTF = UCS transformation format

# Example

Name	Latin A	Hebrew alef	Han AN
Code point	U+0041	+05D0	U+597D
UTF-8	41	D7 90	E5 A5 BD
UTF-16BE	00 41	05 D0	59 7D
UTF-32BE	00 00 00 41	0 00 05 D0	00 00 59 7D

# 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 “%ö”.
2. CCS specifies numeric codes, i.e. ISO 10646 uses 97, 33, 228 and 8240 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.
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.

Note that UTF-8 does not require the shortest possible representation!

0xFE and 0xFF cannot appear in UTF-8.  
0x00 only appears for 0.

# UTF-8: Properties

- ▶ Can encode all  $2^{31}$  UCS characters
- ▶ Characters may take up to 6 octets, `strlen()` will not work!
- ▶ Strings sorted using UCS-4BE will remain sorted
- ▶ Applications **may** reject encodings that use more space than required