Codificação de caracteres

- Representação binária (ou hexa) de cada símbolo
- ASCII (anos 1960)
  American Standard Code for Information Interchange
  - Representação de 128 símbolos (7 bits)
    - 100 0001 = 65 = 0x41 = 'A'
  - O 8º bit (mais significativo) era usado como bit de paridade nas comunicações
  - Faltavam os caracteres latinos acentuados, os chineses, os cirílicos (russos), …

ASCII

0	*null*	16	*data link escape*	32	*space*	48	0	64	@	80	P	96	`	112	p
1	*start of heading*	17	*device control 1*	33	!	49	1	65	A	81	Q	97	a	113	q
2	*start of text*	18	*device control 2*	34	"	50	2	66	B	82	R	98	b	114	r
3	*end of text*	19	*device control 3*	35	#	51	3	67	C	83	S	99	c	115	s
4	*end of transmission*	20	*device control 4*	36	$	52	4	68	D	84	T	100	d	116	t
5	*enquiry*	21	*negative acknowledge*	37	%	53	5	69	E	85	U	101	e	117	u
6	*ackknowledge*	22	*synchoronous idle*	38	&	54	6	70	F	86	V	102	f	118	v
7	*bell*	23	*end of trans. block*	39	'	55	7	71	G	87	W	103	g	119	w
8	*backspace*	24	*cancel*	40	(	56	8	72	H	88	X	104	h	120	x
9	*horizontal tab*	25	*end of medium*	41	)	57	9	73	I	89	Y	105	i	121	y
10	*line feed*	26	*substitute*	42	*	58	:	74	J	90	Z	106	j	122	z
11	*vertical tab*	27	*escape*	43	+	59	;	75	K	91	[	107	k	123	{
12	*form feed*	28	*file separator*	44	,	60	<	76	L	92	\	108	l	124	|
13	*carriage return*	29	*group separator*	45	-	61	=	77	M	93	]	109	m	125	}
14	*shift out*	30	*record separator*	46	.	62	>	78	N	94	^	110	n	126	~
15	*shift in*	31	*unit separator*	47	/	63	?	79	O	95	_	111	o	127	*del*

Codificações de 8 bits
ISO-8859-1 (caracteres latinos)

	_0	_1	_2	_3	_4	_5	_6	_7	_8	_9	_A	_B	_C	_D	_E	_F
8_	U+0080 PAD 128	U+0081 HOP 129	U+0082 BPH 130	U+0083 NBH 131	U+0084 IND 132	U+0085 NEL 133	U+0086 SSA 134	U+0087 ESA 135	U+0088 HTS 136	U+0089 HTJ 137	U+008A LTS 138	U+008B PLD 139	U+008C PLU 140	U+008D RI 141	U+008E SS2 142	U+008F SS3 143
9_	U+0090 DCS 144	U+0091 PU1 145	U+0092 PU2 146	U+0093 STS 147	U+0094 CCH 148	U+0095 MW 149	U+0096 SPA 150	U+0097 EPA 151	U+0098 SOS 152	U+0099 SGCI 153	U+009A SCI 154	U+009B CSI 155	U+009C ST 156	U+009D OSC 157	U+009E PM 158	U+009F APC 159
A_	U+00A0 NBSP 160	U+00A1 ¡ 161	U+00A2 ¢ 162	U+00A3 £ 163	U+00A4 ¤ 164	U+00A5 ¥ 165	U+00A6 ¦ 166	U+00A7 § 167	U+00A8 ¨ 168	U+00A9 © 169	U+00AA ª 170	U+00AB « 171	U+00AC ¬ 172	U+00AD 173	U+00AE ® 174	U+00AF ¯ 175
B_	U+00B0 ° 176	U+00B1 ± 177	U+00B2 ² 178	U+00B3 ³ 179	U+00B4 ´ 180	U+00B5 µ 181	U+00B6 ¶ 182	U+00B7 · 183	U+00B8 ¸ 184	U+00B9 ¹ 185	U+00BA º 186	U+00BB » 187	U+00BC ¼ 188	U+00BD ½ 189	U+00BE ¾ 190	U+00BF ¿ 191
C_	U+00C0 À 192	U+00C1 Á 193	U+00C2 Â 194	U+00C3 Ã 195	U+00C4 Ä 196	U+00C5 Å 197	U+00C6 Æ 198	U+00C7 Ç 199	U+00C8 È 200	U+00C9 É 201	U+00CA Ê 202	U+00CB Ë 203	U+00CC Ì 204	U+00CD Í 205	U+00CE Î 206	U+00CF Ï 207
D_	U+00D0 Ð 208	U+00D1 Ñ 209	U+00D2 Ò 210	U+00D3 Ó 211	U+00D4 Ô 212	U+00D5 Õ 213	U+00D6 Ö 214	U+00D7 × 215	U+00D8 Ø 216	U+00D9 Ù 217	U+00DA Ú 218	U+00DB Û 219	U+00DC Ü 220	U+00DD Ý 221	U+00DE Þ 222	U+00DF ß 223
E_	U+00E0 à 224	U+00E1 á 225	U+00E2 â 226	U+00E3 ã 227	U+00E4 ä 228	U+00E5 å 229	U+00E6 æ 230	U+00E7 ç 231	U+00E8 è 232	U+00E9 é 233	U+00EA ê 234	U+00EB ë 235	U+00EC ì 236	U+00ED í 237	U+00EE î 238	U+00EF ï 239
F_	U+00F0 ð 240	U+00F1 ñ 241	U+00F2 ò 242	U+00F3 ó 243	U+00F4 ô 244	U+00F5 õ 245	U+00F6 ö 246	U+00F7 ÷ 247	U+00F8 ø 248	U+00F9 ù 249	U+00FA ú 250	U+00FB û 251	U+00FC ü 252	U+00FD ý 253	U+00FE þ 254	U+00FF ÿ 255

Codificações de 8 bits
ISO-8859-7 (caracteres gregos)

	_0	_1	_2	_3	_4	_5	_6	_7	_8	_9	_A	_B	_C	_D	_E	_F
8_	U+0080 PAD 128	U+0081 HOP 129	U+0082 BPH 130	U+0083 NBH 131	U+0084 IND 132	U+0085 NEL 133	U+0086 SSA 134	U+0087 ESA 135	U+0088 HTS 136	U+0089 HTJ 137	U+008A LTS 138	U+008B PLD 139	U+008C PLU 140	U+008D RI 141	U+008E SS2 142	U+008F SS3 143
9_	U+0090 DCS 144	U+0091 PU1 145	U+0092 PU2 146	U+0093 STS 147	U+0094 CCH 148	U+0095 MW 149	U+0096 SPA 150	U+0097 EPA 151	U+0098 SOS 152	U+0099 SGCI 153	U+009A SCI 154	U+009B CSI 155	U+009C ST 156	U+009D OSC 157	U+009E PM 158	U+009F APC 159
A_	U+00A0 NBSP 160	U+2018 ‘ 161	U+2019 ’ 162	U+00A3 £ 163	U+20AC € 164	U+20AF ₯ 165	U+00A6 ¦ 166	U+00A7 § 167	U+00A8 ¨ 168	U+00A9 © 169	U+037A ͺ 170	U+00AB « 171	U+00AC ¬ 172	U+00AD 173	174	U+2015 ― 175
B_	U+00B0 ° 176	U+00B1 ± 177	U+00B2 ² 178	U+00B3 ³ 179	U+0384 ΄ 180	U+0385 ΅ 181	U+0386 Ά 182	U+00B7 · 183	U+0388 Έ 184	U+0389 Ή 185	U+038A Ί 186	U+00BB » 187	U+038C Ό 188	U+00BD ½ 189	U+038E Ύ 190	U+038F Ώ 191
C_	U+0390 ΐ 192	U+0391 Α 193	U+0392 Β 194	U+0393 Γ 195	U+0394 Δ 196	U+0395 Ε 197	U+0396 Ζ 198	U+0397 Η 199	U+0398 Θ 200	U+0399 Ι 201	U+039A Κ 202	U+039B Λ 203	U+039C Μ 204	U+039D Ν 205	U+039E Ξ 206	U+039F Ο 207
D_	U+03A0 Π 208	U+03A1 Ρ 209	210	U+03A3 Σ 211	U+03A4 Τ 212	U+03A5 Υ 213	U+03A6 Φ 214	U+03A7 Χ 215	U+03A8 Ψ 216	U+03A9 Ω 217	U+03AA Ϊ 218	U+03AB Ϋ 219	U+03AC ά 220	U+03AD έ 221	U+03AE ή 222	U+03AF ί 223
E_	U+03B0 ΰ 224	U+03B1 α 225	U+03B2 β 226	U+03B3 γ 227	U+03B4 δ 228	U+03B5 ε 229	U+03B6 ζ 230	U+03B7 η 231	U+03B8 θ 232	U+03B9 ι 233	U+03BA κ 234	U+03BB λ 235	U+03BC μ 236	U+03BD ν 237	U+03BE ξ 238	U+03BF ο 239
F_	U+03C0 π 240	U+03C1 ρ 241	U+03C2 ς 242	U+03C3 σ 243	U+03C4 τ 244	U+03C5 υ 245	U+03C6 φ 246	U+03C7 χ 247	U+03C8 ψ 248	U+03C9 ω 249	U+03CA ϊ 250	U+03CB ϋ 251	U+03CC ό 252	U+03CD ύ 253	U+03CE ώ 254	255

Unicode

- Sistema de codificação de caracteres capaz de representar texto de qualquer sistema de escrita existente (mais de 100 mil símbolos)
  - Sistemas latino, arábico, cirílico, chinês, hebráico, …
  - Símbolos matemáticos, geométricos, musicais, setas, ícones, emojis, …
  - Escrita cuneiforme, Braille, runas, élfico, …
  - Próximos: hieróglifos egípcios, alfabeto babilônico, …

- Referências:
  - http://unicode.org/
  - http://www.unicodetables.com/

Unicode

- **UTF-8** (*8-bit Unicode Transformation Format*)
  - Codificação Unicode de comprimento variável (1 a 4 bytes), que pode representar qualquer caráter (*code point*) do padrão.
  - Padrão usado na Internet e na Web
- UTF-16
  - Codificação de 2 ou 4 bytes, usada principalmente para escrita em idiomas dos países asiáticos
- UTF-32
  - Codificação de comprimento fixo de 4 bytes

UTF-8

Bits do *code point*	Primeiro *code point*	Último *code point*	Bytes	Byte 1	Byte 2	Byte 3	Byte 4
7	U+0000	U+007F	1	0xxxxxxx			
11	U+0080	U+07FF	2	110xxxxx	10xxxxxx		
16	U+0800	U+FFFF	3	1110xxxx	10xxxxxx	10xxxxxx	
21	U+10000	U+1FFFFF	4	11110xxx	10xxxxxx	10xxxxxx	10xxxxxx

- 1 byte - Tabela ASCII
- 2 bytes - Caracteres latinos, hebraicos, gregos, …
- 3 bytes - Caracteres chineses, japoneses, coreanos, ...
- 4 bytes - Alguns outros caracteres e símbolos

UTF-8

Representação de strings

C	o	n	c	e	i	ç	ã	o
43	6F	6E	63	65	69	C3A7	C3A3	6F

9 caracteres – 11 bytes