Codificación de Caracteres

Jose Manuel Rodríguez Calvo

Índice

Introducción

La codificación de caracteres no es solamente exclusivo de los ordenadores, sino que también es útil para otras aplicaciones como puede ser el Braille (1824) o el Morse (1837), ambos creados aproximadamente cien años antes de la primera computadora.

En el ámbito de la computación, la codificación es necesaria ya que los ordenadores son sistemas binarios, mientras que el humano es decimal (numérico) y nuestro alfabeto contiene 27 letras (castellano). En las siguientes diapositivas explico los diferentes sistemas de codificación desarrollados desde el Binary Code Decimal (1959) hasta Unicode usado hoy en dia.

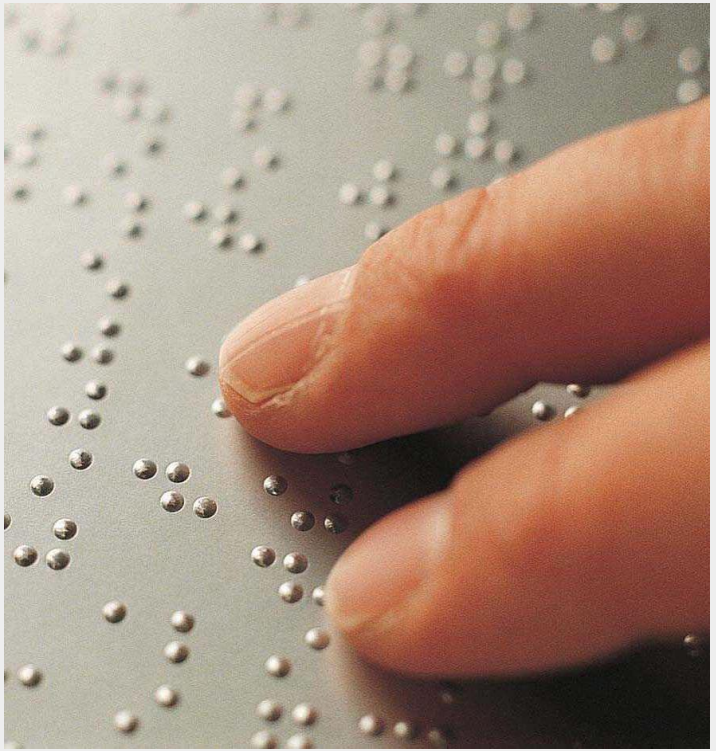


figura 1: codificación de Braille de http://c-pack.com.br/?page=noticias.inst.ler&id=573

Binary Code Decimal

Este sistema de codificación fue uno de los primeros, utilizado en los ordenadores IBM en 1959. Usa seis bits con lo que se puede codificar dígitos en decimal, mayúsculas y algunos signos de puntuación, en total 64 caracteres. Algunos ejemplos de conversión:

- 7= 00 0111
- D= 11 0100

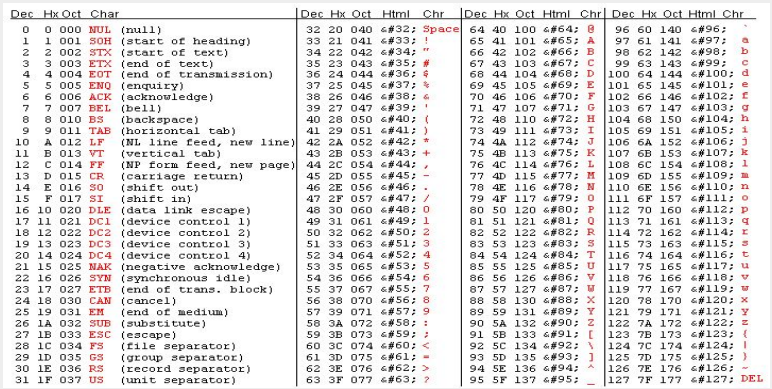
Aunque esto no es suficiente para expresarnos correctamente, debido a la necesidad de expresar minúscula y otros caracteres especiales. Esto da paso al código ASCII

ASCII

Este sistema fue introducido en 1963, y usa una arquitectura de 7 bits, lo que le proporciona 128 caracteres, suficientes para codificar las minúsculas y un gran número de caracteres especiales, aunque esto no era suficiente para codificar otros caracteres especiales de otros idiomas como la ñ o ç .

Extended Binary Coded Decimal Interchange Code

Al igual que el código ASCII este sistema fue desarrollado en 1963 y a diferencia de este usa una arquitectura de 8 bits lo que le da la posibilidad de codificar hasta 256 caracteres.
Mayormente conocido como ASCII extendido es la continuación del código ASCII para expresar caracteres latinos como la ñ (164). En el afán de encontrar un sistema internacional igual para todos nació el estándar Unicode.



Dec	Hx	Oct	Char	Dec	Hx	Oct	Html	Chr	Dec	Hx	Oct	Html	Chr	Dec	Hx	Oct	Html	Chr
0	0	000	NUL (null)	32	20	040		Space	64	40	100	@	@	96	60	140	`	`
1	1	001	SOH (start of heading)	33	21	041	!	!	65	41	101	A	A	97	61	141	a	a
2	2	002	STX (start of text)	34	22	042	"	"	66	42	102	B	B	98	62	142	b	b
3	3	003	ETX (end of text)	35	23	043	#	#	67	43	103	C	C	99	63	143	c	c
4	4	004	EOT (end of transmission)	36	24	044	$	$	68	44	104	D	D	100	64	144	d	d
5	5	005	ENQ (enquiry)	37	25	045	%	%	69	45	105	E	E	101	65	145	e	e
6	6	006	ACK (acknowledge)	38	26	046	&	&	70	46	106	F	F	102	66	146	f	f
7	7	007	BEL (bell)	39	27	047	'	'	71	47	107	G	G	103	67	147	g	g
8	8	010	BS (backspace)	40	28	050	(	(	72	48	110	H	H	104	68	150	h	h
9	9	011	TAB (horizontal tab)	41	29	051	)	)	73	49	111	I	I	105	69	151	i	i
10	A	012	LF (NL line feed, new line)	42	2A	052	*	*	74	4A	112	J	J	106	6A	152	j	j
11	B	013	VT (vertical tab)	43	2B	053	+	+	75	4B	113	K	K	107	6B	153	k	k
12	C	014	FF (NP form feed, new page)	44	2C	054	,	,	76	4C	114	L	L	108	6C	154	l	l
13	D	015	CR (carriage return)	45	2D	055	-	-	77	4D	115	M	M	109	6D	155	m	m
14	E	016	SO (shift out)	46	2E	056	.	.	78	4E	116	N	N	110	6E	156	n	n
15	F	017	SI (shift in)	47	2F	057	/	/	79	4F	117	O	O	111	6F	157	o	o
16	10	020	DLE (data link escape)	48	30	060	0	0	80	50	120	P	P	112	70	160	p	p
17	11	021	DC1 (device control 1)	49	31	061	1	1	81	51	121	Q	Q	113	71	161	q	q
18	12	022	DC2 (device control 2)	50	32	062	2	2	82	52	122	R	R	114	72	162	r	r
19	13	023	DC3 (device control 3)	51	33	063	3	3	83	53	123	S	S	115	73	163	s	s
20	14	024	DC4 (device control 4)	52	34	064	4	4	84	54	124	T	T	116	74	164	t	t
21	15	025	NAK (negative acknowledge)	53	35	065	5	5	85	55	125	U	U	117	75	165	u	u
22	16	026	SYN (synchronous idle)	54	36	066	6	6	86	56	126	V	V	118	76	166	v	v
23	17	027	ETB (end of trans. block)	55	37	067	7	7	87	57	127	W	W	119	77	167	w	w
24	18	030	CAN (cancel)	56	38	070	8	8	88	58	130	X	X	120	78	170	x	x
25	19	031	EM (end of medium)	57	39	071	9	9	89	59	131	Y	Y	121	79	171	y	y
26	1A	032	SUB (substitute)	58	3A	072	:	:	90	5A	132	Z	Z	122	7A	172	z	z
27	1B	033	ESC (escape)	59	3B	073	;	;	91	5B	133	[	[	123	7B	173	{	{
28	1C	034	FS (file separator)	60	3C	074	<	<	92	5C	134	\	\	124	7C	174			|
29	1D	035	GS (group separator)	61	3D	075	=	=	93	5D	135	]	]	125	7D	175	}	}
30	1E	036	RS (record separator)	62	3E	076	>	>	94	5E	136	^	^	126	7E	176	~	~
31	1F	037	US (unit separator)	63	3F	077	?	?	95	5F	137	_	_	127	7F	177	&#127;	DEL

Figura 2: codigo ASCII extendido de http://jbwyatt.com/ascii.html

Unicode

A diferencia de los otros sistemas de codificación Unicode supone un estándar que persigue la universalidad, la uniformidad y unicidad. La importancia de este estándar frente a otros es que marca el esquema para el UTC (Unicode Technical Comitee) del que forman parte un gran número de empresas como Google o Microsoft.

De este estándar podemos desglosar tres formas de codificar UTF-8, UTF-16, UTF-32 (Unicode Transformation Format). El número indica la cantidad de bits que se usan para codificar los mensajes, pudiendo tener 256, 65 536 y hasta 4 294 967 296 caracteres respectivamente.

Cada año se va actualizando la lista de caracteres, donde podemos encontrar que en la última modificación realizada en 2017 se añadieron simbolos como el bitcoin y un grupo de emoticonos hasta alcanzar 136 690 caracteres.

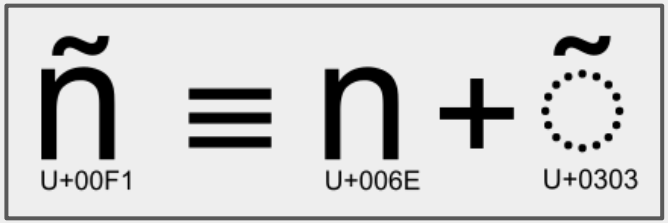


Figura 3: explicación de formacion de caracteres
de https://es.wikipedia.org/wiki/Unicode#/media/File:Composicion_nh.svg

Bibliografía

Codificación de caracteres. (2018, 15 de febrero). *Wikipedia, La enciclopedia libre*. Fecha de consulta: 20:46, septiembre 15, 2018 desde https://es.wikipedia.org/w/index.php?title=Codificaci%C3%B3n_de_caracteres&oldid=105606240.

Wikipedia contributors. (2018, September 13). Character encoding. In *Wikipedia, The Free Encyclopedia*. Retrieved 20:47, September 15, 2018, from https://en.wikipedia.org/w/index.php?title=Character_encoding&oldid=859333155

Wikipedia contributors. (2018, September 15). UTF-8. In *Wikipedia, The Free Encyclopedia*. Retrieved 20:48, September 15, 2018, from https://en.wikipedia.org/w/index.php?title=UTF-8&oldid=859679815

Editors, H. (2018). Morse Code & the Telegraph. Retrieved from https://www.history.com/topics/inventions/telegraph

COMPUTER CODES. (2018). Retrieved 20:51, September 15, 2018, from http://202.114.32.200:8080/courseware/208405/20840511/context/Text/EC2_1.htm

¿Qué es Unicode? (n.d.). Retrieved September 15, 2018, from https://www.softwaredoit.es/definicion/definicion-unicode.html

Unicode. (2018, 25 de agosto). *Wikipedia, La enciclopedia libre*. Fecha de consulta: 21:00, septiembre 15, 2018 desde https://es.wikipedia.org/w/index.php?title=Unicode&oldid=110177732.

Binary Coded Decimal. Recuperado 15 septiembre, 2018, de https://www.electronics-tutorials.ws/binary/binary-coded-decimal.html