CodificaciÃ³n de Caracteres

José Urzúa

jose@nic.cl

NIC Chile

ASCII

- [1963] ASCII: *American Standard Code for Information Interchange*
  - Usa 7 bits para 128 caracteres (0 al 127)
  - 1 byte tiene 8 bits: quedan libres las posiciones entre el 128 y 255
    - Distintos usos e implementaciones
  - Internet: comenzó a compartir textos entre distintos computadores
    - Problemas de compatibilidad

Dec	Hex	Oct	Chr	Dec	Hex	Oct	HTML	Chr	Dec	Hex	Oct	HTML	Chr	Dec	Hex	Oct	HTML	Chr
0	0	000	NULL	32	20	040		Space	64	40	100	@	@	96	60	140	`	`
1	1	001	Start of Header	33	21	041	!	!	65	41	101	A	A	97	61	141	a	a
2	2	002	Start of Text	34	22	042	"	"	66	42	102	B	B	98	62	142	b	b
3	3	003	End of Text	35	23	043	#	#	67	43	103	C	C	99	63	143	c	c
4	4	004	End of Transmission	36	24	044	$	$	68	44	104	D	D	100	64	144	d	d
5	5	005	Enquiry	37	25	045	%	%	69	45	105	E	E	101	65	145	e	e
6	6	006	Acknowledgment	38	26	046	&	&	70	46	106	F	F	102	66	146	f	f
7	7	007	Bell	39	27	047	'	'	71	47	107	G	G	103	67	147	g	g
8	8	010	Backspace	40	28	050	(	(	72	48	110	H	H	104	68	150	h	h
9	9	011	Horizontal Tab	41	29	051	)	)	73	49	111	I	I	105	69	151	i	i
10	A	012	Line feed	42	2A	052	*	*	74	4A	112	J	J	106	6A	152	j	j
11	B	013	Vertical Tab	43	2B	053	+	+	75	4B	113	K	K	107	6B	153	k	k
12	C	014	Form feed	44	2C	054	,	,	76	4C	114	L	L	108	6C	154	l	l
13	D	015	Carriage return	45	2D	055	-	-	77	4D	115	M	M	109	6D	155	m	m
14	E	016	Shift Out	46	2E	056	.	.	78	4E	116	N	N	110	6E	156	n	n
15	F	017	Shift In	47	2F	057	/	/	79	4F	117	O	O	111	6F	157	o	o
16	10	020	Data Link Escape	48	30	060	0	0	80	50	120	P	P	112	70	160	p	p
17	11	021	Device Control 1	49	31	061	1	1	81	51	121	Q	Q	113	71	161	q	q
18	12	022	Device Control 2	50	32	062	2	2	82	52	122	R	R	114	72	162	r	r
19	13	023	Device Control 3	51	33	063	3	3	83	53	123	S	S	115	73	163	s	s
20	14	024	Device Control 4	52	34	064	4	4	84	54	124	T	T	116	74	164	t	t
21	15	025	Negative Ack.	53	35	065	5	5	85	55	125	U	U	117	75	165	u	u
22	16	026	Synchronous idle	54	36	066	6	6	86	56	126	V	V	118	76	166	v	v
23	17	027	End of Trans. Block	55	37	067	7	7	87	57	127	W	W	119	77	167	w	w
24	18	030	Cancel	56	38	070	8	8	88	58	130	X	X	120	78	170	x	x
25	19	031	End of Medium	57	39	071	9	9	89	59	131	Y	Y	121	79	171	y	y
26	1A	032	Substitute	58	3A	072	:	:	90	5A	132	Z	Z	122	7A	172	z	z
27	1B	033	Escape	59	3B	073	;	;	91	5B	133	[	[	123	7B	173	{	{
28	1C	034	File Separator	60	3C	074	<	<	92	5C	134	\	\	124	7C	174			|
29	1D	035	Group Separator	61	3D	075	=	=	93	5D	135	]	]	125	7D	175	}	}
30	1E	036	Record Separator	62	3E	076	>	>	94	5E	136	^	^	126	7E	176	~	~
31	1F	037	Unit Separator	63	3F	077	?	?	95	5F	137	_	_	127	7F	177	&#127;	Del

Basados en ASCII

- [1987] ISO-8859-1 o Latin1
  - Incluye 191 caracteres
    - Basado en conjunto de caracteres de terminales VT220
    - Considera vocales con tilde, letra ñ
  - Ampliamente usado en América, Europa Occidental, Oceanía y parte de África
  - Es la base de conjuntos de caracteres populares de 8-bit, incluyendo Windows-1252

Unicode

- [1991] Unicode:
  - Conjunto de caracteres que pretende incluir todos los caracteres de uso común
  - Provee las bases para procesar, almacenar e intercambiar texto en cualquier lenguaje
  - Define un “code point” para cada caracter
  - V10.0 en junio/2017, incluye 136.690 caracteres (56 emoticonos)

A. Summary

Unicode 10.0 adds 8,518 characters, for a total of 136,690 characters. These additions include 4 new scripts, for a total of 139 scripts, as well as 56 new emoji characters.

The new scripts and characters in Version 10.0 add support for lesser-used languages and unique written requirements worldwide, including:

- Masaram Gondi, used to write Gondi in Central and Southeast India
- Nüshu, used by women in China to write poetry and other discourses until the late twentieth century
- Soyombo and Zanabazar Square, used in historic Buddhist texts to write Sanskrit, Tibetan, and Mongolian
- Syriac letters used for writing Suriyani Malayalam, also known as Garshuni and as Syriac Malayalam
- Gujarati signs used for the transliteration of the Arabic script into Gujarati by Ismaili Khoja communities
- A set of 285 Hentaigana characters used in Japan (historic variants of Hiragana characters)
- CJK Extension F (7,473 Han characters)

Important symbol additions include:

- Bitcoin sign
- 56 emoji characters (full list)
- A set of Typicon marks and symbols

For statistics regarding emoji associated with Unicode 10.0, see Emoji Counts.

Unicode

- rongorongo

Fuente: https://imaginaisladepascua.com/la_isla_de_pascua/cultura_rapa_nui/rongorongo/

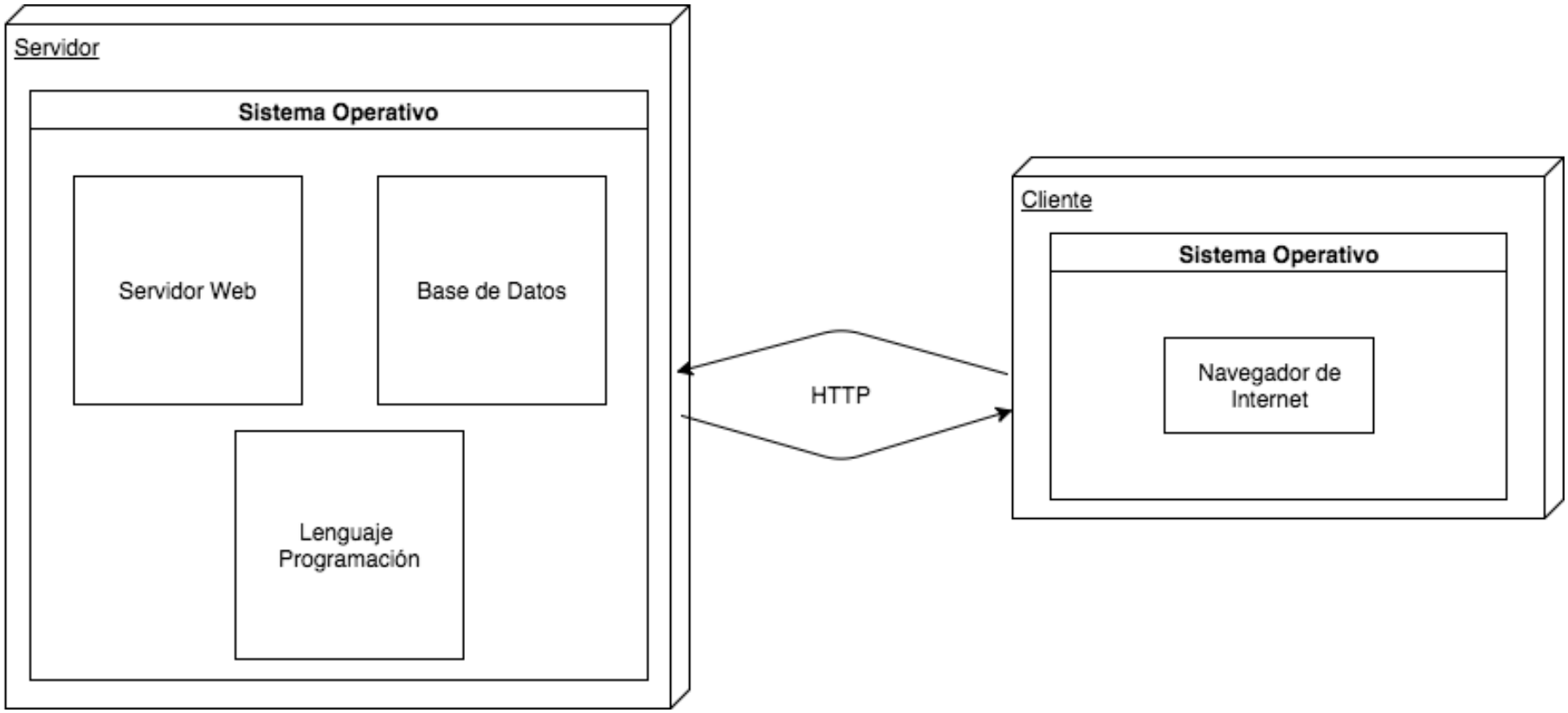
UTF-8

- UTF: Unicode Transformation Format
  - Algoritmo para asignar a cada caracter de Unicode una secuencia única de bytes
  - UTF-8 es el más común para aplicaciones web
    - Codificación de ancho variable: 1, 2, 3 o 4 bytes
  - Coincide con tabla ASCII en las posiciones 32 a 127
    - Letras, números, símbolos
    - Compatible con codificación ASCII

Recordar

¡Texto plano no existe!

Aplicación web



Aplicación Web

- Sistema operativo: revisar “locale”

```
[~ jourzua]$ locale
LANG="en_US.UTF-8"
LC_COLLATE="en_US.UTF-8"
LC_CTYPE="en_US.UTF-8"
LC_MESSAGES="en_US.UTF-8"
LC_MONETARY="en_US.UTF-8"
LC_NUMERIC="en_US.UTF-8"
LC_TIME="en_US.UTF-8"
LC_ALL="en_US.UTF-8"
[~ jourzua]$ locale charmap
UTF-8
```

Aplicación Web: Base de Datos

- MySQL: Revisar configuración de servidor

```
[mysqld]
character-set-server=utf8
collation-server=utf8_general_ci
```

- Revisar charset disponibles
- Indicar charset al momento de crear una base de datos.
  - También al crear tablas y columnas

Aplicación Web: Base de Datos

```
mysql>  SHOW CHARACTER SET like 'utf%';
+---------+-------------------+--------------------+--------+
| Charset | Description       | Default collation  | Maxlen |
+---------+-------------------+--------------------+--------+
| utf8    | UTF-8 Unicode     | utf8_general_ci    |      3 |
| utf8mb4 | UTF-8 Unicode     | utf8mb4_general_ci |      4 |
| utf16   | UTF-16 Unicode    | utf16_general_ci   |      4 |
| utf16le | UTF-16LE Unicode  | utf16le_general_ci |      4 |
| utf32   | UTF-32 Unicode    | utf32_general_ci   |      4 |
+---------+-------------------+--------------------+--------+
5 rows in set (0.00 sec)

mysql> CREATE DATABASE ejemplo_utf8 CHARACTER SET utf8;
Query OK, 1 row affected (0.00 sec)
```

Aplicación Web: Servidor web

- Protocolo HTTP 1.1:
  - Configuración cabecera “Content-Type”
  - Documento web el default charset es ISO-8859-1
  - Apache: ver configuración “AddDefaultCharset”

AddDefaultCharset Directive

Description:	Default charset parameter to be added when a response content-type is `text/plain` or `text/html`
Syntax:	`AddDefaultCharset On|Off|`*`charset`*
Default:	`AddDefaultCharset Off`
Context:	server config, virtual host, directory, .htaccess
Override:	FileInfo
Status:	Core
Module:	core

Aplicación Web: Lenguaje Programación

- Entorno desarrollo, codificación de código fuente y archivos de texto, caracter BOM
- URL de conexión a base de datos

```
jdbc:mysql://localhost:3306/mydb?
        useUnicode=true&amp;characterEncoding=utf8
```

- Considerar TAG <meta> en cabecera HTML, en los primeros 1024 bytes:

```
<meta charset="utf-8”>
```

Recomendaciones

- Conversión entre codificaciones: `iconv`

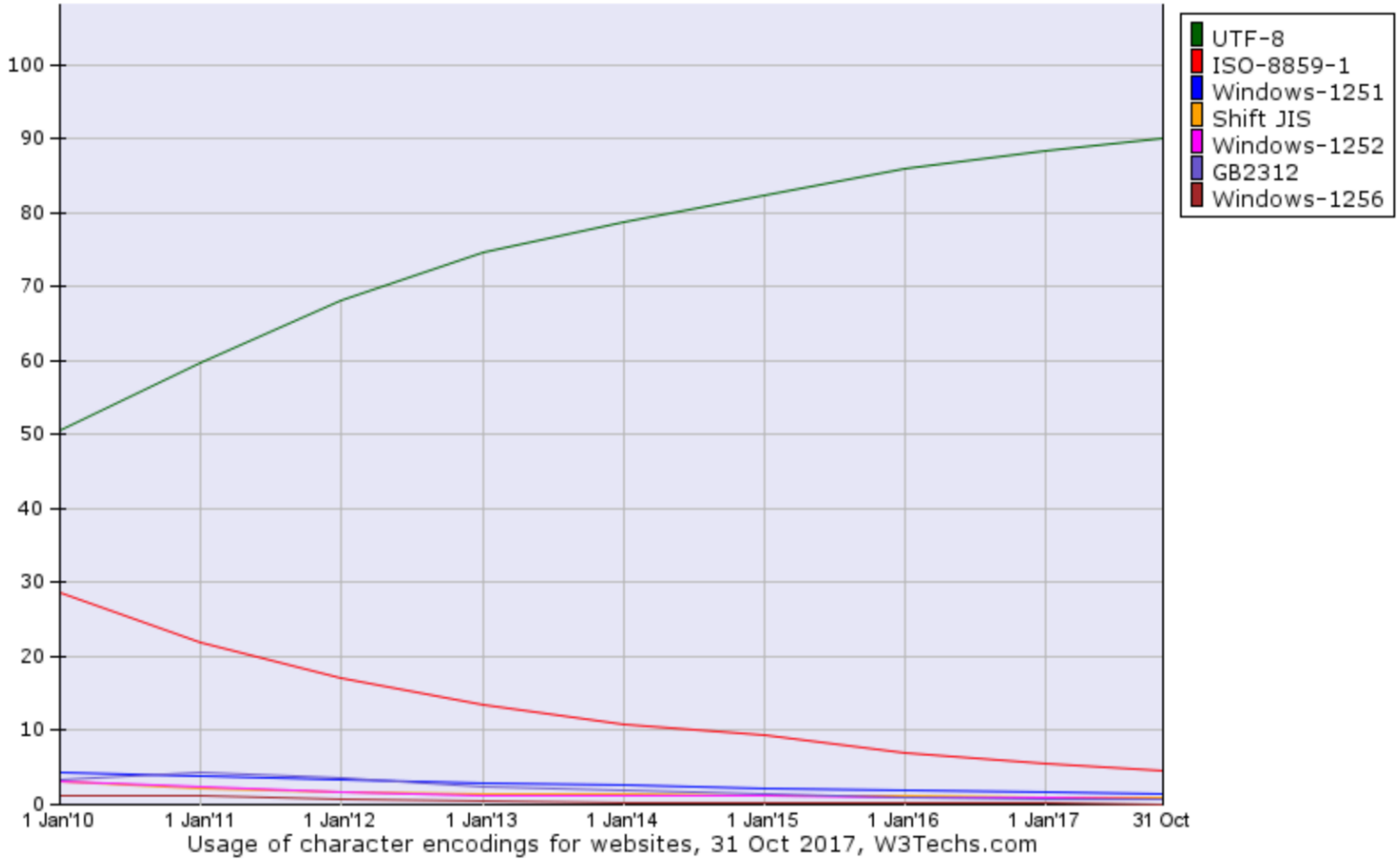
```
NAME
       iconv - character set conversion

SYNOPSIS
       iconv [OPTION...] [-f encoding] [-t encoding] [inputfile ...]
       iconv -l

DESCRIPTION
       The iconv program converts text from one encoding to another encod-
       ing.  More precisely, it converts from the encoding given  for  the
       -f  option to the encoding given for the -t option. Either of these
       encodings defaults to the encoding of the current locale.  All  the
       inputfiles  are  read  and  converted  in  turn; if no inputfile is
       given, the standard input is used. The converted text is printed to
       standard output.
```

Recomendaciones

- ¿Qué encoding se usa? UTF-8!



Recomendaciones

- ¿Qué dice W3C?

QUICK ANSWER

Choose UTF-8 for all content and consider converting any content in legacy encodings to UTF-8.

Fuente: https://www.w3.org/International/questions/qa-choosing-encodings#quickanswer

CodificaciÃ³n de Caracteres

José Urzúa
jose@nic.cl
NIC Chile

@jourzua