



Escuela de
Ingeniería y Arquitectura
Universidad Zaragoza



Departamento de
Informática e Ingeniería
de Sistemas
Universidad Zaragoza



Tema 3: Juegos de caracteres

Garantía y Seguridad (30242)

especialidad en

4º curso, Grado en Ingeniería Informática

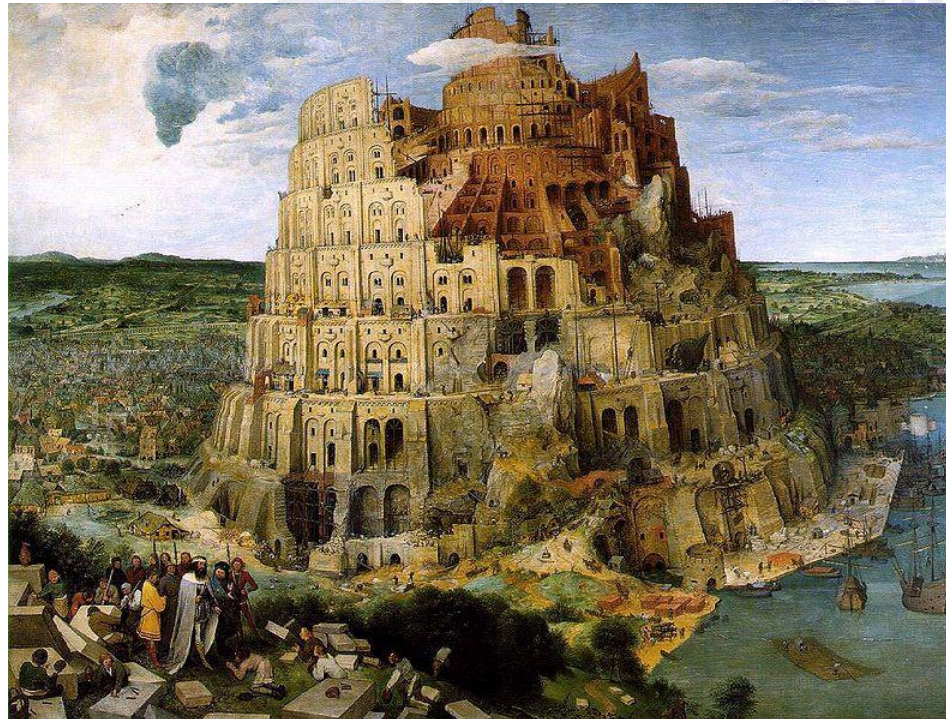
Ingeniería de Computadores

J. Alastruey, P. Ibáñez, V. Viñals

Área de Arquitectura y Tecnología de Computadores (ATC)

Juegos de caracteres

Definiciones
ASCII, ISO10646, Unicode



Representación de cadenas de caracteres

- ❑ Problema: recepción de secuencia de bytes
 - ¿Cómo se traduce a caracteres?

41 72 61 67 F3 6E 0D 0A
↓
A r a g ó n \n

Representación de cadenas de caracteres

❑ ¿Qué caracteres?

- a, e, i, o, u
- á, é, í, ó, ú

Repertorio

❑ Muchas tablas de mapeo

- ASCII
- Familia ISO Latin

Código de carácter

❑ x octetos = 1 carácter

Codificación

Definiciones

- ❑ **Repertorio de caracteres:** conjunto de caracteres (concepto)
 - Nombre y representación (glyph: logotipo)
- ❑ **Código de carácter:** correspondencia entre caracteres de un repertorio y números naturales (mapeo por tabla)
- ❑ **Codificación de carácter:** método (algoritmo) para convertir números de código en secuencia de octetos
- ❑ Ejemplo:

disco de ideograma
Festos CJK

Repertorio	a	!	ä	%	∞		儻
Códigos ISO10646 (hex)	61	21	E4	2030	221E	101D0	513B
Codificación 2 bytes/carácter	00 61	00 21	00 E4	20 30	22 1E	np	np

Ejemplos de códigos de carácter: ASCII

□ Repertorio

SP ! " # \$ % & ' () * + , - . / 0 1 2 3 4 5 6 7 8 9 : ; < = > ? @
A B C D E F G H I J K L M N O P Q R S T U V W X Y Z [\] ^ _
` a b c d e f g h i j k l m n o p q r s t u v w x y z { | } ~

□ Códigos

32, 33 ...	64
65, 66 ...	95
96, 97 ...	126

□ Codificación

Ejemplos de códigos de carácter

❑ ASCII

- US-ASCII: 95 caracteres + 33 control = 128 (1 octeto/carácter)
- Variantes nacionales (`#$@[\]^_`{|}~`)
- 8-bit ASCII = extensión de ASCII

❑ ISO 646: similar a US-ASCII, pero reserva varios códigos para “caracteres nacionales”: `@[\]{|}`

❑ ISO 8859-1 (ISO Latin 1)

- 190 + control = 256
- 1 octeto/carácter

❑ Windows code page 1252 (WinLatin1)

- Códigos 128-159: caracteres (ISO Latin1 = control)
- Distinto del juego de caracteres DOS ¡!
- Por defecto en MySQL

Códigos de carácter: familia ISO 8859

- ❑ ISO 8859-1: Latin-1, mayoría de lenguas del Oeste de Europa (Danés, Español, Holandés, Inglés, Faroés, Finlandés, Francés, Alemán, Islandés, Irlandés, Italiano...), alguno del Este (Albanés), Indonesio y las africanas Afrikaans y Swahili
- ❑ ISO 8859-2: Latin-2, lenguas de Europa Central y del Este que usan el alfabeto latino (Bosnio, Polaco, Croata, Checo, Eslovaco, Esloveno, Serbio, Húngaro)
- ❑ ISO 8859-3: Latin-3, Sur de Europa (Turco, Maltés y Esperanto)
- ❑ ISO 8859-4: Latin-4, Norte de Europa (Estonio, Letón, Lituano, Groenlandés, Sami)
- ❑ ISO 8859-5: Latin/Cyrilic, Lenguas eslavas que usan alfabeto Cirílico (Bielorruso, Búlgaro, Macedonio, Ruso, Serbio) y Ucraniano
- ❑ ISO 8859-6: Latin/Arabic, caracteres más comunes del árabe
- ❑ ISO 8859-7: Latin/Greek, griego moderno
- ❑ ISO 8859-8: Latin/Hebrew, alfabeto hebreo usado en Israel
- ❑ ISO 8859-9: Latin-5, turco (8859-1 adaptado al turco)
- ❑ ISO 8859-10: Latin-6, lenguas nórdicas. Las bálticas usan más 8859-4
- ❑ ISO 8859-11: Latin/Thai
- ❑ ISO 8859-12: Latin/Devanagari, abandonada
- ❑ ISO 8859-13: Latin-7, Baltic Rim
- ❑ ISO 8859-14: Latin-8, Celtas (Gaélico, Bretón)
- ❑ ISO-8859-15: Latin-9, revisión de 8859-1, contiene €
- ❑ ISO-8859-16: Latin-10, Albanés, Croata, Húngaro, Italiano, Polaco, Rumano, Esloveno, contiene €

Ejemplos de códigos de carácter

- ❑ DOS: CP codepages
- ❑ Macintosh
- ❑ Hasta aquí para todos:
 - ASCII es subconjunto
 - Codificación con 8 bits
 - Todos menos Win1252: posiciones 128-159 son códigos de control
- ❑ Pero hay más ...
 - EBCDIC: mainframes IBM
- ❑ Problemas
 - Interoperabilidad
 - Intercambio entre distintos idiomas
 - Limitación caracteres que pueden mostrarse en un documento

¿solución?

- ❑ Define Universal multiple-octet coded Character Set (UCS)
 - Todos los caracteres de todas las lenguas escritas
 - Símbolos matemáticos, signos, ligaduras
 - ...
- ❑ ~100.000 caracteres identificados por
 - Nombre no ambiguo
 - Código (code point)
- ❑ Espaciado proporcional
- ❑ Logotipos ≠ Caracteres
- ❑ Independientes de la lengua
- ❑ Caracteres base y combinantes
 - Compuestos

$$A + ^{\circ} = \text{Å}$$
$$U+0041 + U+030A = U+00C5$$

ISO 10646-Unicode

- ❑ Consorcio Unicode sincronizado desde 1991 con ISO 10646
 - Adobe, Apple, Google, IBM, Microsoft, Oracle, Yahoo
- ❑ Versiones
 - Unicode 1.0.0 (oct-1991): 7.161 caracteres
 - Unicode 1.0.1 (jun-1992): 28.359 caracteres
 - ...
 - Unicode 6.3 (sep-2013): 110.122 caracteres
 - Unicode 8.0 (jun-2015): 120.737 caracteres
 - ◆ 5 modificadores tono piel
 - Unicode 9.0 (jun-2016): 128.172 caracteres
 - ◆ http://www.unicode.org/emoji/charts/emoji-versions.html#v9.0_2016
 - ◆ <http://www.unicode.org/reports/tr51/tr51-7.html>
 - Unicode 10.0 (jun-2017): 136.690 caracteres
 - Unicode 11.0 (jun-2018): 137.374 caracteres
 - Unicode 12.0 (mar-2019): 137.993 caracteres
 - ◆ 61 nuevos emojis
 - Unicode 13.0 (mar-2020): 143.859 caracteres
 - Unicode 14.0 (sep-2021):

ISO 10646-Unicode

❑ Representación código: U+nnnn (hex)

- U+0020: space
- U+0041: latin capital letter A
- U+1F4A9: pile of poo



Ubuntu: Ctrl+Shift+u
Ej: 7F (126, ~)

❑ Caracteres clasificados en 17 planos

- 0 (0000–FFFF): Basic Multilingual Plane (BMP)
- 1 (10000–1FFFF): Supplementary Multilingual Plane (SMP)
- 2 (20000–2FFFF): Supplementary Ideographic Plane (SIP)
- 3-13 (30000–DFFFF): sin asignar
- 14 (E0000–EFFFF): Supplementary Special-purpose Plane (SSP)
- 15 (F0000–FFFFF): reservado (Private Use Area, PUA)
- 16 (100000–10FFFF): reservado (Private Use Area, PUA)

Basic Multilingual Plane (BMP)

0000	US-ASCII, Latinos, Cirílico, Griego, Hebreo, Índico
1000	y otros alfabetos con "pocos" caracteres
2000	Símbolos: puntuación, científicos, matemáticos,...
3000	CJK (Chino, Japonés, Koreano):
4000	Caracteres fonéticos, símbolos y signos puntuac.
5000	
6000	Ideogramas CJK (20.902)
7000	
8000	
9000	
A000	
B000	Sílabas "Hangul" Koreano (11.712)
C000	
D000	
E000	Otros ...
F000	

Basic Multilingual Plane (BMP)

00	01	02	03	04	05	06	07	08	09	0A	0B	0C	0D	0E	0F
10	11	12	13	14	15	16	17	18	19	1A	1B	1C	1D	1E	1F
20	21	22	23	24	25	26	27	28	29	2A	2B	2C	2D	2E	2F
30	31	32	33	34	35	36	37	38	39	3A	3B	3C	3D	3E	3F
40	41	42	43	44	45	46	47	48	49	4A	4B	4C	4D	4E	4F
50	51	52	53	54	55	56	57	58	59	5A	5B	5C	5D	5E	5F
60	61	62	63	64	65	66	67	68	69	6A	6B	6C	6D	6E	6F
70	71	72	73	74	75	76	77	78	79	7A	7B	7C	7D	7E	7F
80	81	82	83	84	85	86	87	88	89	8A	8B	8C	8D	8E	8F
90	91	92	93	94	95	96	97	98	99	9A	9B	9C	9D	9E	9F
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

- Latin scripts and symbols
- Linguistic scripts
- Other European scripts
- Middle Eastern and Southwest Asian scripts
- African scripts
- South Asian scripts
- Southeast Asian scripts
- East Asian scripts
- Unified CJK Han
- Canadian Aboriginal scripts
- Symbols
- Diacritics
- UTF-16 surrogates and private use
- Miscellaneous characters
- Unallocated code points

Codificaciones para ISO 10646-Unicode

- ❑ Unicode Transformation Format (UTF)
 - Transformación código Unicode a secuencia única de bytes
 - Repertorio ASCII queda tal cual
- ❑ UTF-7 (7 bits) (para smtp ...)
 - Códigos <128 ->1 byte
 - Resto códigos: n bytes (0-127)
- ❑ UTF-8 (8 bits) (rfc 2279)
 - Códigos <128 ->1 byte
 - Resto códigos: 2-6 bytes (0-255)
- ❑ UTF-16, UTF-16BE, UTF-16LE (UCS-2)
 - Códigos codificados en secuencias de 2 bytes
- ❑ UTF-32, UTF-32BE, UTF-32LE (UCS-4)
 - 4 bytes por carácter (única codificación de longitud fija)

```
◆Subject: T+AOQ-m+AOQ-  
◆Content-Type: text/plain;  
charset=UTF-7  
◆Content-Transfer-Encoding: 7bit  
◆Catal+-AOA-, Fran+-AOc-ais
```

```
◆Subject: =?UTF-8?B?VMokbcOk?=  
◆Content-Type: text/plain;  
charset=UTF-8  
◆Content-Transfer-Encoding: 8bit  
◆CatalÃ, FranÃ§ais
```

Unicode Transformation Format (UTF)

❑ Propiedades

Name	UTF-8	UTF-16	UTF-16BE	UTF-16LE	UTF-32	UTF-32BE	UTF-32LE
Smallest code point	0000	0000	0000	0000	0000	0000	0000
Largest code point	10FFFF	10FFFF	10FFFF	10FFFF	10FFFF	10FFFF	10FFFF
Code unit size	8 bits	16 bits	16 bits	16 bits	32 bits	32 bits	32 bits
Byte order	N/A	<BOM>	big-endian	little-endian	<BOM>	big-endian	little-endian
Fewest bytes per character	1	2	2	2	4	4	4
Most bytes per character	4	4	4	4	4	4	4

<BOM> indicates that the byte order is determined by a byte order mark, if present at the beginning of the data stream, otherwise it is big-endian.

Bytes	Encoding form
00 00 FE FF	UTF-32, big-endian
FF FE 00 00	UTF-32, little-endian
FE FF	UTF-16, big-endian
FF FE	UTF-16, little-endian
EF BB BF	UTF-8

Codificaciones para ISO 10646-Unicode

❑ UTF-8: RFC 3629

- <https://www.rfc-editor.org/info/rfc3629>

UCS-4 (hex)	UTF-8 (bin)
0000 0000 - 0000 007F	0xxxxxxx
0000 0080 - 0000 07FF	110xxxxx 10xxxxxx
0000 0800 - 0000 FFFF	1110xxxx 10xxxxxx 10xxxxxx
0001 0000 - 001F FFFF	11110xxx 10xxxxxx
0020 0000 - 03FF FFFF	111110xx 10xxxxxx
0400 0000 - 7FFF FFFF	1111110x 10xxxxxx

	e	ñ	e
Hexa	0x65	0xF1	0x65
Bin	0110 0101	1111 0001	0110 0101
UTF-8		110xxxxx 10xxxxxx	

Ejemplos

- ❑ UTF-8: mayoría de distribuciones linux
- ❑ UTF-16: Windows/NTFS, Java
- ❑ UTF-32: linux, Unix

- ❑ Conversiones
 - iconv (Linux)
 - uconv (UNIX)

```
$ locale
LANG=es_ES.UTF-8
LANGUAGE=
LC_CTYPE="es_ES.UTF-8"
LC_NUMERIC="es_ES.UTF-8"
LC_TIME="es_ES.UTF-8"
LC_COLLATE="es_ES.UTF-8"
LC_MONETARY="es_ES.UTF-8"
LC_MESSAGES="es_ES.UTF-8"
LC_PAPER="es_ES.UTF-8"
LC_NAME="es_ES.UTF-8"
LC_ADDRESS="es_ES.UTF-8"
LC_TELEPHONE="es_ES.UTF-8"
LC_MEASUREMENT="es_ES.UTF-8"
LC_IDENTIFICATION="es_ES.UTF-8"
LC_ALL=
```

Juego de Caracteres	Repr. (hex) de "Ä"
ISO Latin 1	C4
Unicode (ISO 10646), UTF-16	000000C4
Unicode, codificación UTF-8	E3C4
PC850 (MSDOS)	8E
MacOS	80

Ejercicios

entes términos (repertorio, código, ...):

- ISO 10646
- US-ASCII
- CP850
- WinLatin1 (W1252).

ficación de un texto con UTF-8 de Unicode/ISO
de usar ISO Latin-1 (8859-1) con 8 bits/carácter.
ño total en bytes, extensión del juego de caracteres,
.

- # Ejercicios
-
- entes términos (repertorio, código, ...):
- ISO 10646
 - US-ASCII
 - CP850
 - WinLatin1 (W1252).
- ficación de un texto con UTF-8 de Unicode/ISO
de usar ISO Latin-1 (8859-1) con 8 bits/carácter.
ño total en bytes, extensión del juego de caracteres,
.

Bibliografía

- ❑ Juegos de caracteres:
 - <http://xfiles.rainmaker.iki.fi/jkorpela/chars.html>
 - <http://crl.nmsu.edu/Resources/presentations/ogden/ogden-sls/>
- ❑ Unicode
 - <http://www.unicode.org/>