2.4 ESTÁNDAR UNICODE

Por ***Alberto Prieto Espinosa***
Profesor Emérito del Departamento de Arquitectura y Tecnología de los Computadores de la UGR

Inconvenientes de los códigos tradicionales (SBCD, EBCDIC, ASCII, etc.)

- **Los símbolos codificados son insuficientes para representar los caracteres especiales que requieren numerosas aplicaciones.**
- **Los símbolos y códigos añadidos en las versiones ampliadas a 8 bits no están normalizados.**
- **Están basados en los caracteres latinos, existiendo otras culturas que utilizan otros símbolos muy distintos.**
  - Los lenguajes escritos de diversas culturas orientales, como la china, japonesa y coreana se basan en la utilización de ideogramas o símbolos que representan palabras, frases o ideas completas, siendo, por tanto, inoperantes los códigos que sólo codifican letras individuales.

19

Unicode (ISO/IEC 10646)

- **Propuesto en por un consorcio de empresas y entidades que trata de hacer posible escribir aplicaciones que sean capaces de procesar texto de muy diversas culturas. Se busca**
  - Universalidad,
    - trata de cubrir la mayoría de lenguajes escritos existentes en la actualidad: Inicialmente **16 bits ⇒ 65.356 símbolos** (ASCII ampliado: 256 caracteres)
  - Unicidad,
    - a cada carácter se le asigna exactamente un único código (idiogramas con imagen distinta, tienen igual código), y
  - Uniformidad,
    - ya que ***inicialmente*** todos los símbolos se representan con un número fijo de bits (***16***).

20

Asignación de posiciones (*puntos de código*) en el Plano Básico Multilingüe (*BPM*)

Zona	*Códigos*		*Símbolos codificados*	*Nº de caractere*
A	0000	0000 00FF	Latín-1	256
			otros alfabetos	7.936
		2000	Símbolos generales y caracteres fonéticos chinos, japoneses y coreanos	8.192
I	4000		Ideogramas	24.576
O	A000		Pendiente de asignación	16.384
R	E000 FFFF		Caracteres locales y propios de los usuarios. Compatibilidad con otros códigos	8.192

21

Subjonjuntos Unicode estándarizados

Rango Unicode	*Se corresponde con*
0000 a 007F	Latín Básico (00 a 7F), definidos en la norma ASCII ANSI-X3.4.
0080 a 00FF	Suplemento Latín-1 (ISO 8859-1)
0100 a 017F	Ampliación A de Latín
0180 a 024F	Ampliación B del Latín
0250 a 02AF	Ampliación del Alfabeto Fonético Internacional (IPA)
02BF a 02FF	Espaciado de letras modificadoras
0300 a 036F	Combinación de marcas diacríticas (tilde, acento grave, etc.)
0370 a 03FF	Griego
0400 a 04FF	Cirílico
0530 a 058F	Armenio
0590 a 05FF	Hebreo
0600 a 06FF	Árabe
0700 a 074F	Sirio
etc.	etc.

22

Con el tiempo se han ido realizando ampliaciones, incluyendo nuevos "planos"

- En el BPM hay asignados sólo 24.576 puntos de código para ideogramas. El diccionario de la RAE contiene unas 88.000 palabras; pero una persona no suele utilizar más de unas 11.000.
- En la actualidad (Unicode 5.2 , 2009) hay asignados o reservados 17 planos → 17 x 216 = 1.114.112 puntos de código dentro del rango de 0000 a 10FFFF .
- En general, un punto Unicode se referencia escribiendo "U+" seguido por su nº HEX.

Plano 0	Basic Multilingual Plane (BMP)	0000–FFFF
Plano 1:	Supplementary Multilingual Plane (SMP):	10000–1FFFF
Plano 2	Supplementary Ideographic Plane (SIP):	20000–2FFFF
Plano 3–13	Sin asignar	30000–DFFFF
Plano 14:	Supplementary Special-purpose Plane (SSP)	E0000–EFFFF
Planos 15–16	Supplementary Private Use Area (S PUA A/B)	F0000–10FFFF

23