Unicode

Introducción

Este artículo explica como representa texto el ordenador. Desde porque los ordenadores usan el sistema binario, qué es ASCII, a conceptos generales de Unicode. Unicode es un estandar complejo, pero aquí tienes la cultura general necesaria para hablar del tema, y orientarte por ti mismo.

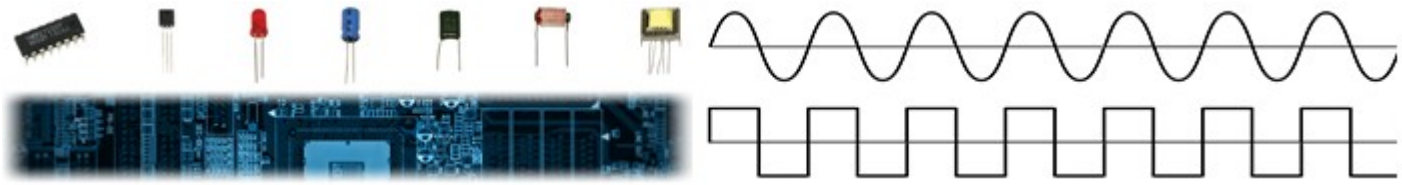
El sistema binario

Un **componente electrónico** es un dispositivo capaz de manipular electrones o sus campos asociados. Algunos ejemplos son transistores, diodos, y condensadores.

Los componentes electrónicos se combinan y conectan mediante cables para formar **circuitos electrónicos** capaces de realizar operaciones complejas. La radio es un circuito electrónico que transforma una señal electromagnética en una señal sonora.

A un circuito electrónico lo llamamos **analógico** si opera con rangos continuos de señales electricas. Por ejemplo, un altavoz amplifica la señal a un número indefinido de valores entre un volumen mínimo y máximo.

En cambio, si opera con señales discretas, decimos que el circuito electrónico es **digital**. Un ordenador es un circuito digital que opera con dos voltajes de valores exactos: cero voltios, y el número de voltios que proporcione la fuente de voltaje, por ejemplo 1.4 voltios. Esto es así porque están construidos con **transistores**, que son componentes electrónicos cuyo funcionamiento es más fiable cuando representa solo dos estados. A grandes rangos, cuantos más estados representas con un rango finito de voltaje, más posibilidades existen de confundir un estado con el otro.



En resumen, los ordenadores expresan información usando solo dos estados porque así sus componentes son más fiables.

El sistema de numeración más fácil de implementar con dos estados es el **sistema numérico binario**. El binario es similar al decimal, pero solo utiliza los digitos uno y cero. Por ejemplo:

decimal	0	1	2	3	4	…
binario	0	1	10	11	100	…

Bits y bytes

Cuando transferimos texto usando un ordenador, existen componentes electrónicos y cables que leen y escriben estas cadenas de ceros y unos. Cada uno de los digitos 0 o 1 que circulan por el ordenador, se considera la unidad mínima de información en informática, y se conoce como **bit**.

Para acelerar la transferencia, los bits se transfieren en grupos de 8, 16, 32, o 64. Cuantos más bits lees a la vez, más cables y componentes necesitas para transferirlos. El equilibrio entre miniaturización, coste de construcción, y otras consideraciones prácticas, hace que los ordenadores actuales operen con bloques de 64 bits. Sin embargo hace 50 años, el límite práctico estaba en 8. Un grupo de 8 bits se conoce como **byte**.

Las unidades de información son multiplos de dos por dos motivos:

- El ordenador usa el sistema de numeración binario tanto para expresar datos, como para referenciar la posición en memoria de esos datos.
- El numero de valores expresable con un digito es base$^{\text{número de digitos}}$ (que es un multiplo de la base).

En consecuencia, la capacidad expresiva de un digito es completamente aprovechada si referencia una cantidad de información que multiplo de su base. Por ejemplo con un digito decimal (base 10) puedes expresar 10 valores (0, 1, 2, 3, 4, 5, 6, 7, 8, 9). Por tanto si usaramos la base decimal, lo lo más eficiente sería referenciar información en multiplos de diez, de otro modo desperdiciriamos la capacidad expresiva de un dígito.

El motivo por el que además son multiplos de 8 es historico. Han existido computadores en los que un byte tenía 1, 6, 7, 8, 9, 12, 18, y 20 bits, pero la popularidad del chip 8080 (el cual usaba 8-bit) hizo que byte fuera sinónimo de 8 bits.

¿Qué es un caracter?

Un **caracter** es una unidad de información abstracta. Algunos ejemplos de caracteres son los números y letras del alfabeto, los signos de puntuación, los emojis, y los caracteres de control como un salto de línea.

Un **juego de caracteres** es una colección de caracteres usada para escribir en un lenguaje particular. Algunos idiomas usan uno, otros usan varios (por ejemplo, kanji y kana del japones), y otros usan variantes de un mismo juego de caracteres (por ejemplo, todos los lenguajes europeos usan variantes del alfabeto romano).

Un **glifo** es una representación visual de un caracter. En un ordenador, los glifos asociados a caracteres suelen estar almacenados en **fuentes**, también llamados tipos de letra. Cada fuente contiene un estilo homogeneo de glifos para un alfabeto. Este es un ejemplo de varios glifos que representan un mismo caracter:

Tipos de caracteres

A grandes rasgos hay tres sistemas de escritura

- **Alfabético**. Un alfabeto es el conjunto de letras usadas para escribir un lenguaje. Cada una representa un fonema o se usa junto a otras para representar un fonema. La mayoría de alfabetos en uso se basan en el alfabeto romano.

- **Silábico**. Contiene caracteres que representan cada una de las sílabas de un lenguaje. Una sílaba es una combinación de vocal o vocal y consonante. Es apropiado para lenguajes con pocas silabas, como el japones, que contiene un centenar. El ingés en cambio, contiene miles de combinaciones de conjuntos de vocales y consonantes.
- **Ideográfico**. Contiene caracteres que representan ideas. Por ejemplo, el chino, o el egipcio antiguo.

Estas categorías no son rígidas. Por ejemplo, el español es alfabético, pero también usa un ideograma como el símbolo de euro, que representa la estabilidad de Europa (es una e de Europa con dos líneas indican estabilidad).

ASCII

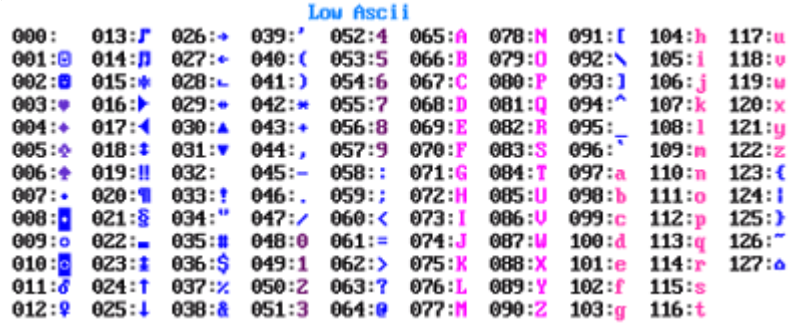
¿Qué es ASCII?

El único tipo de dato que un ordenador es capaz de almacenar y manipular es el número binario. El truco para representar texto es asignar un número a cada letra del alfabeto, e implementar un sistema de renderizado que visualiza el caracter correspondiente a cada número.

Llamamos **código** al sistema de normas para convertir información simple en información codificada. Una **codificación de caracteres** por ejemplo, es un código que convierte caracteres a números.

La codificación de caracteres más popular se llama **ASCII** (American Standard Code for Information Interchange). ASCII asigna números decimales del 0 al 127 a las letras del alfabeto inglés, los números, y algunos signos de puntuación y de control. Los signos de control son indicadores para procesar información, por ejemplo, el salto de línea. La letra “A” en ASCII tiene asignado el número 65 (`0x41` en hexadecimal), y las letras consecutivas del alfabeto tienen asignados números consecutivos:

letra	binario	decimal
A	1000001	65
B	1000010	66
C	1000011	67
D	1000100	68
etc.		



Según este código, la representación de las letras “abc” en binario es la siguiente:

```
Shell


1000001 1000010 1000011

1 1000001 1000010 1000011
```

El documento que estas leyendo en este momento tiene una representación en binario en tu ordenador en forma de ceros y unos. A nivel físico se representa en la memoria mediante la presencia o ausencia de voltaje, y en el disco mediante magnetismo.

Cuando ASCII se publicó en 1964, las máquinas de entonces transferian 8 bits a la vez. Sin embargo ASCII usó solo 7, y reservó el octavo bit para detectar errores de transmisión. Esos 7 bits permiten expresar $2^7 = 128$ valores, que ASCII uso para codificar 33 señales de control (usadas entre máquinas), y 95 números, símbolos de puntuación, y letras del alfabeto inglés.

Variantes del ASCII

En los ’80 aparecieron variantes de ASCII que usaban el octavo bit para representar caracteres adicionales, lo que permitio representar caracteres adicionales propios de alfabetos no ingleses, como por ejemplo el español.

Debido a que 8 bits no son suficientes para representar todos los alfabetos del mundo, continuaron apareciendo variantes ASCII de 8 bits incompatibles entre sí. Estas variantes se llaman a veces **ASCII extendido**, pero no forman parte del estandar ANSI.

Hay varios conjuntos de “ASCII extendido”, cada uno de los cuales contiene codificaciones para muchos lenguajes:

- **Windows code pages**, usado en aplicaciones gráficas Windows.
- **OEM code pages**, usando en aplicaciones de consola Windows.
- **ISO-8859** es un estandar ISO para codificación en 8 bits. Tiene 16 partes. La primera se llama ISO-8859-1, también conocida como Latin-1, que cubre la mayoría de lenguajes de Europa occidental.

En ASCII solo es posible trabajar con un alfabeto a la vez. ASCII tampoco es válido para representar alfabetos asiáticos, porque contienen miles de caracteres. Estas deficiencias, junto a la popularidad de la informática, y la potencia creciente de los ordenadores, hicieron deseable y posible la creación de un estandar más exhaustivo.

Unicode

¿Qué es Unicode?

Unicode es una codificación de caracteres que asigna un número a cada uno de los caracteres de practicamente todos los alfabetos existentes, incluyendo las lenguas muertas como el egipcio antiguo y otros. El estandar Unicode lo publica y mantiene el consorcio Unicode, una entidad sin animo de lucro formada por empresas e individuos.

El estandar Unicode también especifica algoritmos sobre como manejar texto, que probablemente solo necesites si realizas tareas especializadas. Por ejemplo:

- Dividir palabras y líneas.
- Ordenar texto.
- Formatear números, fechas, y horas.
- Mostrar texto que fluye de derecha a izquierda.
- Mostrar texto cuya forma escrita se combina y reordena.
- Tratar problemas de seguridad relativos a caracteres parecidos.

Unidades y puntos de código

Unicode comenzó como un sistema de codificación de 16 bits, aunque desde Unicode 2.0 (publicado en 1996), es una codificación de 21 bits.

Un **punto de código** es el número con el que se identifica un caracter en el estandar Unicode. El punto de código se escribe con el formato `U+xxxx` donde las `xxxx` son de cuatro a seis digitos en sistema de numeración hexadecimal.

En Unicode 7.0 el rango válido de puntos de código va de 0 a $10FFFF_{16}$. El hexadecimal se usa por conveniencia en lugar del binario, porque es más fácil recordar. Por ejemplo, Unicode asigna el número 65 a la letra a latina mayúscula. El punto de código correspondiente es `U+0041` porque $65_{\text{decimal}} = 0x41_{\text{hexadecimal}}$.

Unicode es compatible con la codificación ISO-8859-1, porque los 256 primeros caracteres de Unicode coinciden con los caracteres de ISO-8859-1. Esto hace que la mayoría de texto en uso requiera solo un byte por caracter.

El punto de código se representa con grupos de 8, 16, o 32 bits dependiendo respectivamente, de si el tipo de codificación es UTF-8, UTF-16, o UTF-32. Cada uno de estos grupos se llama unidad de código. Una **unidad de código** es el mínimo grupo de bits necesario para representar una unidad de texto codificado. En UTF-8 es 8 bit, en UTF-16 es 16 bit, y en UTF-32 es 32 bit. Por ejemplo, `U+0041` requiere 2 bytes, es decir, dos unidades de código en UTF-8, y una unidad de código en UTF-16.

Unicode no dice nada sobre como debe representarse cada caracter. Por ejemplo, cada uno de los tipos de letra instalados en tu ordenador contiene una representación visualmente diferente de la letra latina A.

Estructura de Unicode

Planos

Unicode se divide en **17 planos**, que son grupos con 65,536 caracteres cada uno. Por tanto los caracteres representables son 17* (2^{16}) = 1,114,112. Sin emabargo, de esos 17 planos, solo 6 tienen caracteres asignados. Y en esos planos, solo hay 112,956 caracteres asignados en Unicode 7.0.

La mayoría de caracteres usados en lenguas occidentales están en el plano 0, que se conoce como **Plano Multilingue Básico**, también conocido por su acrónimo inglés “BMP”.

Plano	Rango	Nombre
Plano 0	`0x0000` a `0xFFFF`	Plano Multilingue Básico (BMP)
Plano 1	`0x10000` a `0x1FFFF`	Plano Multilingue Suplementario
Plano 2	`0x20000` a `0x2FFFF`	Plano Suplementario Ideográfico
Plano 3 a 13	no asignado	
Plano 14	`0xE0000` a `0xEFFFF`	Plano Suplementario de Proposito Especial
Plano 15	`0x0F0000` a `0x10FFFF`	Plano Suplementario de Uso Privado

En el BMP, los 256 primeros caracteres de Unicode se representan con un byte, mientras que el resto de caracteres necesita dos bytes, porque es el número de bits necesario para abarcar el rango de valores del primer plano (2^{16} = 65536).

Bloques

- Un **bloque** es un rango continuo de puntos de código.
- Cada bloque tiene un nombre único. Por ejemplo “Basic Latin”, “Hebreo”, y otros.
- Cada punto de código tiene una propiedad «Block name» que indica el bloque al que pertenece.
- Los bloques de código pueden consultarse en estos enlaces:
    - Unicode Character Code Charts muestra nombres de bloques, nombres de caracteres, y puntos de código
    - Unicode/Character reference muestra rangos y planos
    - Unicode block (wikipedia)

Emoji

- Los **emoji** son pictogramas (caracteres que representan imagenes).
- La palabra emoji viene del Japones 絵 (e ≅ imagen) 文 (mo ≅ escritura) 字 (ji ≅ caracter).
- Los emoji representan cosas variadas, como emociones, actividades, banderas, vehículos, comida, etc.
- Un **emoticono** (del inglés emoción e icono) es un modo de mostrar una emoción mediante caracteres. Algunos de los emoji son emoticonos porque representan emociones.
- Los emoji están esparcidos en varios bloques. Las expresiones faciales por ejemplo, están en el rango de puntos de código `1F600–1F64F`.
- En plataformas Apple, los glifos para emoji están en la fuente “Apple Color Emoji”.

Para mostrar la paleta de caracteres Emoji en OS X pulsa dos veces ^⌘ + espacio. Si pulsas botón derecho sobre un emoji y pulsas “Copy Character Info”, aparece la siguiente información:

```
Shell


    GRINNING FACE Unicode: U+1F600 (U+D83D U+DE00), UTF-8: F0 9F 98 80
```

```
        GRINNING FACE
1    Unicode: U+1F600
2 (U+D83D U+DE00),
  UTF-8: F0 9F 98 80
```

¿Qué quieren decir cuando dicen…?

Carácter compuesto

Un **carácter compuesto** es una entidad Unicode que puede definirse como una secuencia de otros caracteres.

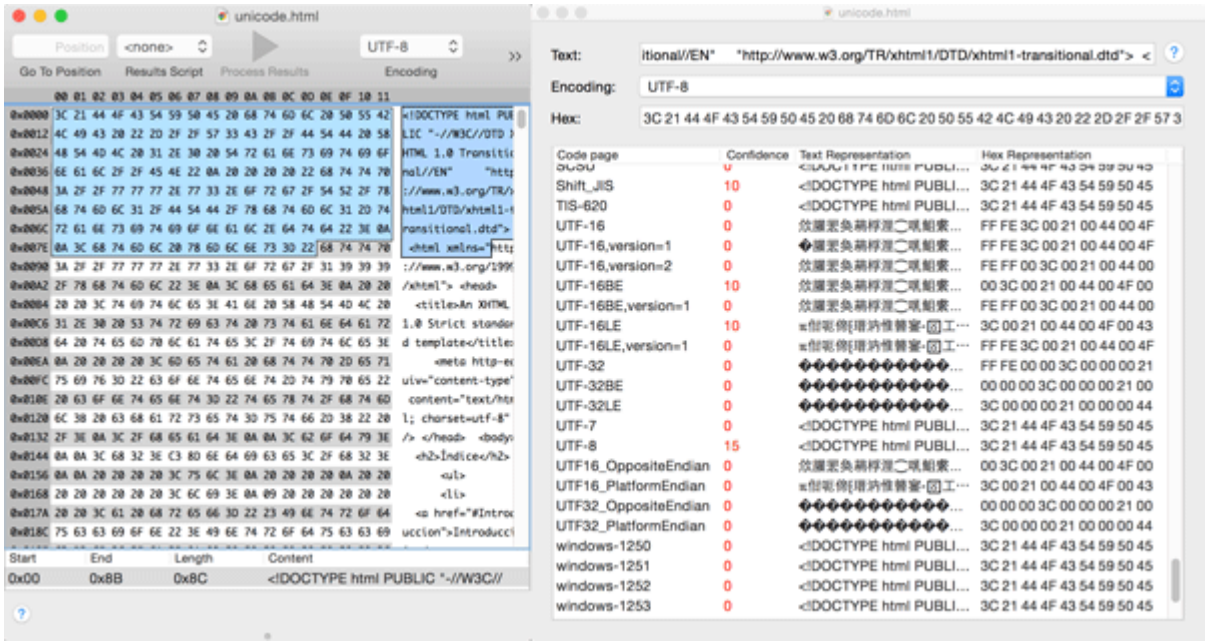
- Por ejemplo, `U+00E9` (letra minúscula latina e con acento agudo) tiene el mismo significado y apariencia que `U+0065 U+0301` (letra minúscula latina e, y acento agudo de combinación). Se consideran **canónicamente equivalentes**, pero no iguales porque están hechos de diferentes puntos de código.
- Los caracteres compuestos existen para permitir implementaciones de Unicode en sistemas que solo permiten representar cada caracter por separado.

Los **caracteres compatibles equivalentes** son aquellos que representan el mismo caracter abstracto, pero tienen diferentes apariencias visuales. Por ejemplo, el carácter ﬀ (ligadura latina minúscula `ff`, `U+FB00`) es compatible con (pero no canónicamente equivalente a) la secuencia `ff` (dos letras latinas minúsculas f, `U+0066 U+0066`).

Si queremos comparar texto, por ejemplo durante una busqueda, quizas queramos tratar caracteres canónicamente equivalentes, y caracteres compatibles equivalentes como un mismo caracter. En ese caso necesitaremos aplicar un **algoritmo de normalización** que convierten una cadena a una representación única que puede ser comparada binariamente. Hay cuatro formas de normalización llamadas C, D, KD, y KC. Estos algoritmos permiten comparar caracteres, pero pueden eliminar distinciones de formato que evitan conversiones en ambos sentidos.

En resumen,

- En **tamaño**, UTF-8 es óptimo para caracteres latinos, UTF-16 para caracteres no latinos, y UTF-32 desperdicia memoria en casi todos los casos.
- En **procesamiento** necesario para obtener el punto de código, UTF-32 es óptimo (no necesita procesamiento). UTF-8 y UTF-16 sí porque porque el número de unidades de código por punto es variable.
- En **transmisión**, las unidades de código de UTF-16 y UTF-32 suelen usar el orden de bytes (endianness) de la CPU en la que se ejecuta una implementación dada. Esto solo es relevante cuando cada palabra tiene más de un byte, así que en UTF-8 no importa porque cada unidad de código solo tiene un byte.



Lenguajes de programación

ANSI y ISO-8859-1 siguen presentes en software creado cuando no existía Unicode o librerias maduras para procesarlo. Más raramente, también se usan porque es más rápido de procesar, o lo bastante bueno para comunicar información básica en software que no está de cara al usuario.

UTF-8 es con gran diferencia la codificación más usada. Los caracteres que coinciden con ISO-8859-1 usan un byte por caracter, el resto dos bytes, y casi siempre que encontremos más de dos, se tratará de un caracter emoji, o de caracteres japoneses, chinos, o koreanos.

Al programar delegaremos las complicaciones de Unicode a las librerías del sistema, pero debemos tener presente estos casos:

- Distinguir entre número de caracteres y número de bytes.
- Contar caracteres en cadenas que soportan caracteres combinados, pares sustitutos, y secuencias de variación.
- Comparar cadenas normalizadas para dar como iguales formas variantes de un mismo caracter.
- Nivel de soporte de Unicode en el motor de expresiones regulares. Hay tres niveles.
- Twitter por ejemplo, cuenta caracteres independientemente de los bytes necesarios para representarlos.