# Representación interna de caracteres. Estándares ASCII y Unicode

Sabemos que los ordenadores trabajan con números utilizando la notación binaria para representarlos No obstante, ¿cómo trabajan con caracteres (es decir, letras minúsculas y mayúsculas, signos de puntuación, etc)? Es decir, ¿cómo reconocen cada caracter que se pulsa en una tecla y cómo saben qué caracter hay que representar en cada momento en una pantalla?

La respuesta es que existe un determinado número de caracteres predefinidos de fábrica en el hardware de toda máquina moderna que ya están listos para poder ser utilizados por cualquier sistema o programa. Todos estos caracteres se representan con una combinación de 7 bits estandarizada por la organización ANSI (American Nacional Standards Institute) en un documento llamado llamado ASCII (American Standard Code for Interchange – Código estándar para el intercambio de información). Este documento en realidad es una tabla de correspondencia (grabada, tal como hemos dicho, en la placa base de cualquier máquina del mundo) que asocia una determinada combinación de 7 bits a un determinado caracter. Esta tabla se muestra aquí: http://www.asciitable.com o también se puede consultar, en Linux, mediante el comando *ascii* (con la posibilidad de indicar el parámetro *-b* para ver la equivalencia de cada caracter en binario, o *-x* para verla solo en hexadecimal o *-d* para verla solo en decimal)

> NOTA: Las 32 primeras combinaciones del código ASCII, utilizadas originalmente para funciones de control de dispositivos se denominan “códigos de control” y son códigos no imprimibles (como el tabulador,, el salto de línea o el retroceso), aunque la mayoría está ya en desuso.

Cuando se trabaja con datos, no obstante, es más rápido y más eficaz para la CPU utilizar 8 bits –un byte- en lugar de utilizar 7 bits. Como resultado, cuando se almacena un dato utilizando la codificación ASCII, ocupa ocho bits, igual que cualquier otro dato, pero con la particularidad que en el código ASCII el bit octavo es “off”. Algunos fabricantes de hardware aprovecharon la posibilidad de disponer de otras 128 combinaciones con el octavo bit “on” y crearon lo que se conoce generalmente como conjunto de caracteres "ASCII extendido”, los cuales incluyen símbolos internacionales de monedas, caracteres para dibujar líneas, alfabetos distintos del inglés y otros caracteres que pueden imprimirse. No obstante, este conjunto "extendido" de caracteres no era estrictamente ASCII (es decir, estandarizado), por lo que variaban según el creador, y, por tanto, al final crearon varios problemas de compatibilidad entre máquinas.. El más extendido de ellos fue el llamado conjunto ISO-8859-1.

Además, de los problemas de compatibilidad, pronto se vio la necesidad de adoptar un estándar de representación de caracteres que fuera más amplio que el ASCII extendido ya que con éste igualmente sólo se tenían $2^8$ caracteres diferentes posibles (recordemos la fórmula de contar combinaciones), quedando excluida totalmente la representación de caracteres chinos, árabes, hebreos, cirílicos, etc. Surgió entonces el estándar Unicode (http://www.unicode.org) realizado por un consorcio de empresas. La gran diferencia entre ASCII y Unicode es que este último es un estándar software ; esto significa que el estándar ASCII sigue estando presente de fábrica en todos los dispositivos electrónicos actuales pero en el momento de instalar un sistema operativo sobre él, se instalarán unas determinadas tablas Unicode correspondientes al idioma elegido. Por lo tanto, en un sistema con el idioma español habrá instaladas ciertas tablas Unicode y en otro sistema con el idioma coreano habrá instaladas otras tablas Unicode. De esta forma, ya no hay limitación en el número de caracteres reconocidos ni en el de tablas utilizables por un sistema determinado porque todo es software instalable/desinstalable.

Dentro del estándar Unicode hay, de todas formas, diferentes formatos de representación: el más habitual es el denominado UTF-8, en el cual se utilizan al menos 1 byte para la representación de los caracteres más habituales (como pasaba con el "ASCII extendido") pero que en algunos casos pueden utilizarse 2 (o en casos muy raros hasta 3 o 4) ; otro formato común es el UTF-16, el cual se utiliza siempre 2 bytes para la representación de los caracteres, sea cual sea. De esta manera, en UTF-8 se tienen $2^8$ combinaciones posibles (si se usa 1 byte) y $2^{16}$ combinaciones (si es necesario usar 2 bytes para caracteres más "exóticos"), ¡suficientes para todos los alfabetos del mundo!. Por otro lado, el estándar Unicode es compatible con el ASCII; es decir, los 256 primeros caracteres de Unicode coinciden con los del ASCII extendido más popular, (el ISO-8859-1) de manera que hay problemas de conversión con aplicaciones más antiguas.

Para saber el código Unicode asociado a un determinado carácter se pueden consultar las tablas más populares en http://unicode-table.com . Para saber si un determinado carácter Unicode es reconocido por nuestro sistema (porque se haya instalado el paquete de idioma correspondiente) en el escritorio Gnome se puede utilizar el programa Gucharmap.

**Gestió de les "locales":**

Una "locale" és una configuració regional definida pel sistema que inclou principalment el llenguatge amb què es mostraran els missatges a pantalla i el mapejat del teclat però també el format dels números, hores, dates, telèfons, monedes, unitats de mesura, etc. Aquesta configuració es global per tots els programes/serveis/usuaris el sistema, encara que cada programa/servei/usuari en particular podria tenir la seva pròpia configuració regional definida a les seves preferències; en aquest cas la configuració particular sobreeescriuria la global. Per esbrinar la "locale" actualment definida es pot executar la comanda *localectl [status]*

...la qual simplement mostra el contingut dels arxius "/etc/locale.conf" i "/etc/vconsole.conf". El que es veu representen les diferents variables d'entorn que defineixen cadascuna un determinat aspecte de la configuració regional (el que hem comentat: format de números, dates, llenguatge, etc) i la variable LANG, la qual representa la configuració genèrica. Per conèixer els valors i formats concrets que admet una determinada variable es pot executar *locale -ck LC_NOMVARIABLE* o bé consultar la pàgina del manual *man 5 locale* (per conèixer el significat de cada variable cal consultar la pàgina *man 7 locale*). Per veure totes les locales instal.lades al sistema es pot fer: *localectl list-locales*

Per instal.lar una nova "locale" (per exemple, el francès, "fr_FR", a Ubuntu es pot executar la comanda *check-language-support -l fr_FR* ; aquesta comanda mostra tots els paquets que caldrien instal.lar per disposar d'aquesta "locale" plenament incorporada al sistema. A Fedora només cal instal.lar el metapaquet adient a la "locale" desitjada, el qual té el nom de "langpacks-XX" (on "XX" representa la "locale" escollida)

Per establir temporalment (mentre el terminal actual estigui obert) una "locale" només cal fer *export LANG=fr_FR.utf8* Per establir permanentment una "locale" (gravant-se a l'arxiu "/etc/locale.conf") en canvi cal fer el següent (el canvi es veurà al següent inici de sessió): *localectl set-locale LANG=en_GB.utf8* (o LC_*=...)

> NOTA: Es pot modificar a mà l'arxiu ~/.config/locale.conf per sobreescriure aquesta configuració general per una particular de l'usuari en qüestió.

> NOTA: També és possible indicar una eventual "locale" particular en un determinat arranc del sistema mitjançant paràmetres del kernel. Concretament amb el paràmetre l*ocale.LANG=* i/o els paràmetres més específics (*locale.LC_NUMERIC=, locale.LC_TIME=, locale.LC_MONETARY=, locale.LC_MESSAGES=, locale.LC_PAPER=, locale.LC_ADDRESS=, locale.LC_TELEPHONE=, locale.LC_MEASUREMENT=*, etc, etc)

Per veure tots els mapejats de teclat disponibles es pot fer *localectl list-keymaps*

Per establir permanentment un mapejat de teclat (gravant-se a l'arxiu "/etc/vconsole.conf" ) es pot fer *localectl set-keymap en_GB*

> NOTA: Existeix el paràmetre del kernel vconsole.keymap=... que permet sobreescriure durant l'arranc del sistema la configuració escrita a "/etc/vconsole.conf"

Cal aclarir que la comanda anterior només funciona en terminals virtuals degut a què en entorns gràfics la configuració del teclat ve determinada per la indicada al panell de control de l'escriptori.

La comanda *localectl* en realitat no és més que un client D-Bus que realitza les peticions mitjançant aquesta via (ja siguin de consulta o d'actualització de dades) al servei *systemd-localed*, el qual es posa en marxa només quan és rep la petició (no està tota l'estona funcionant).