Laboratorio 4. Representación de la información

Objetivos:

- Experimentar con distintos formatos de codificación de textos y aprender a averiguar el tipo de contenido de un fichero examinando las cabeceras internas.
- Conocer el formato y practicar el uso de herramientas que permiten cambiar metadatos, y manipular ficheros de audio mp3.

Recursos:

- El ordenador del laboratorio o cualquier equipo con un sistema de tipo Unix.
- Portal Moodle de la asignatura.

Resultados:

Como resultado de esta práctica deben subirse a Moodle dos ficheros (respetando las reglas sobre el nombre de dichos ficheros que se indican en el apartado 4):

- Un formulario (resultados-lab4-2018.odt [1]disponible en Moodle) que debe ir rellenando mientras realiza la práctica, siguiendo las indicaciones **en negrita** incluidas en diversos apartados de este enunciado.
- Un fichero de audio mp3, obtenido a partir de las indicaciones que se describen en el apartado 3 de este documento.

Actividades previas:

Antes de la sesión de laboratorio, debe leer previamente este documento e intentar realizar las actividades indicadas para asegurar que puede completarlo antes de finalizar la sesión.

1. Codificación de textos

Empiece desde el entorno gráfico, abra un terminal de texto, cree un nuevo directorio `lab4`, vaya a este directorio `lab4` y baje del Moodle el fichero `resultados-lab4-2018.odt`.

Sin moverse del directorio `lab4` realice paso a paso las siguientes actividades:

A. Cree un fichero con el nombre `apell` que contenga los caracteres de su primer apellido (no ponga tildes y, si tiene alguna ñ, cámbiela por n), seguidos de un espacio y el símbolo €. Puede hacerlo con un editor de texto o, de forma más sencilla, usando la orden

```
$ echo su-apellido € > apell
```
[2]

Para comprobar la creación del fichero `apell` visualice su contenido en la pantalla.

B. Utilice la orden `wc -c` para contar el número de bytes del fichero `apell`.

Tenga en cuenta para los siguientes apartados que los editores de texto y la orden `echo` añaden el carácter `0x0A` (salto de línea) al final de cada línea.

---

[1] La extensión *.odt* se emplea para ficheros de documentos creados con la suite ofimática gratuita ***LibreOffice*** (http://www.libreoffice.org). Para editar el fichero, puede usar LibreOffice o, si lo prefiere, MS Office Word.

[2] En lo sucesivo el símbolo `$` significa que debe escribir la orden indicada y, si en la descripción de la orden aparece algún campo en cursiva, como *su-apellido*, debe sustituirlo por un valor concreto (su primer apellido).

A la vista del número de bytes y teniendo en cuenta el contenido del fichero, ¿puede deducir si está codificado en ISO Latin9 (ISO 8859-15) o en UTF-8? **Anote en el formulario el número de bytes del fichero y responda a la pregunta formulada, justificando su respuesta.**

C. Compruebe la codificación del fichero `apell` con la orden `file`:

```
$ file apell
```

D. El programa `iconv` convierte un fichero de una codificación a otra. La orden

```
iconv -f codificación-de-fich -t codificación-result fich
```

convierte el contenido del fichero `fich`, codificado como se indica a continuación de `-f` (from) a la codificación que se indica a continuación de `-t` (to) y da el resultado por la salida estándar.

Se pueden ver todas las codificaciones posibles con la orden `iconv -l`. Dos de ellas son: LATIN9 (equivalente a ISO 8859-15) y UTF-8[3].

Ahora, si su fichero `apell` está en UTF-8, conviértalo a LATIN9 y guárdelo en otro fichero de nombre `apell-cod` con la órden:

```
$ iconv -f UTF-8 -t LATIN9 apell > apell-cod
```

Si está en LATIN9 invierta los valores de los parámetros `-f` y `-t` para convertirlo a un fichero codificado en UTF-8.

Con la orden `file` compruebe que ha cambiado la codificación del fichero `apell-cod`

E. Mire con `wc -c` el número de bytes del nuevo fichero `apell-cod`. **Copie al formulario la salida de las órdenes `file apell` y `file apell-cod` y explique las diferencias que se aprecian en ambos ficheros.**

F. El programa `hd` (“*hexadecimal dump*”) permite ver los contenidos binarios (representados en hexadecimal) de un fichero, mostrando en la salida estándar tres columnas: en la primera columna se indican las direcciones relativas de los bytes (empezando por 0x0), en la segunda columna se muestra la expresión en hexadecimal de cada byte (16 bytes en cada línea) y en la tercera columna se muestra su posible interpretación como ASCII (si no la tiene, muestra un punto).

Compruebe con la orden `hd` los contenidos de los dos ficheros (`apell` y `apell-cod`), observando la correspondencia entre caracteres y codificaciones en ambos ficheros y analizando sus diferencias. **Copie al formulario la salida que genera `hd` para ambos ficheros `apell` y `apell-cod` y explique sus diferencias.**

G. Los ficheros que se han manipulado en los anteriores apartados son ficheros de texto “plano”. Pero, cuando se edita un fichero con un *procesador de texto* se generan “ficheros enriquecidos” que, además de los caracteres, incluyen propiedades (color, tamaño, tipo de letra…). Dos extensiones estándares son bien conocidas: `.odt` (usada en OpenOffice) y `.docx` (en Microsoft Word) que realmente corresponden a *archivos*, es decir a ficheros conteniendo a su vez un conjunto de directorios y ficheros de texto (en xml) y que están comprimidos con ZIP.

[3] Puede comprobar cuál es la codificación por defecto en su sistema con echo $LANG

Para ver qué ficheros y directorios contiene el *archivo* `resultados-Lab4-2018.odt` (el formulario de entrega que está rellenando) ejecute la orden:

```
$ unzip -l resultados-lab4-2018.odt
```

Copie al formulario el resultado que aparece en pantalla.

2. Identificación del tipo de contenidos: Metadatos

Para el sistema de ficheros todos los ficheros regulares son iguales, pero los programas que trabajan con ellos necesitan conocer qué es lo que contienen: texto, imagen, o código ejecutable. Hay varias maneras de hacerlo y una de ellas es incluir datos sobre el tipo de contenido (es decir, *metadatos*) dentro del propio fichero, normalmente al principio delante de los datos propiamente dichos. Para explorar cómo identificar el tipo de contenidos de un fichero, realice los siguientes pasos:

A. Copie en su directorio `lab4` un fichero que contenga una imagen codificada con el formato GIF, poniéndole como nombre `imag.gif`. Esta imagen GIF puede descargarla de la web, pero en su sistema hay cientos de imágenes que puede localizar con la orden `locate .gif`

Primero mire qué tipo de contenido tiene el fichero `imag.gif` con la orden:

```
$ file imag.gif
```

Luego copie el fichero `imag.gif` a un fichero `imag.exe`

```
$ cp imag.gif imag.exe
```

y compruebe, con la orden `file` si el fichero `imag.exe` es de tipo ejecutable. **Escriba en el formulario el resultado de la comprobación y añada una explicación sobre la relación que observa entre el tipo de un fichero y su extensión.**

Cambie los permisos de fichero `imag.exe`, con `chmod`, de forma que pueda ser ejecutado por el propietario, el grupo y el resto de usuarios. Debe mantener el permiso de lectura para que `file` pueda leer la cabecera.

Compruebe, con la orden `file` si el fichero `imag.exe` es de tipo ejecutable. **Escriba en el formulario el resultado de la comprobación y añada una explicación sobre la relación entre el tipo de un fichero y sus permisos.**

B. La cabecera de un fichero de tipo GIF empieza con la "firma" (3 bytes) que representan las codificaciones ASCII de los caracteres "GIF", y sigue con la "versión" (3 bytes), que puede ser la codificación de los caracteres "87a" o la de los caracteres "89a"[4].

Con la orden `hd` puede comprobar la información de cabecera del fichero `imag.gif` pero, como el listado resulta ser muy largo, para ver sólo las primeras líneas, escriba la orden

```
$ hd imag.gif | less
```

Copie al formulario los primeros 20 dígitos hexadecimales de la primera línea. De ellos, empiece por observar los 12 primeros y busque en la tabla de representación de caracteres ASCII a qué 6 caracteres corresponden (si no tiene la tabla, haga `man ascii`).

[4] La especificación completa está en http://www.w3.org/Graphics/GIF/spec-gif89a.txt

Copie al formulario los 6 caracteres resultantes. Observe luego los 4 dígitos hexadecimales siguientes e interprételos como un número entero codificado en formato de coma fija de 16 bits, almacenado con convenio extremista menor. **Copie al formulario el número entero resultante**. Repita lo mismo con los 4 dígitos hexadecimales siguientes. **Copie este segundo número al formulario y explique qué relación guardan los números obtenidos con el resultado que muestra la orden `file imag.gif`.**

C. Genere un fichero de texto plano de nombre `raro.gif` que contenga solamente "GIF89a2018" (respetando mayúsculas y minúsculas) y compruebe el tipo con:

```
$ file raro.gif
```

Explique en el formulario por qué el resultado de la orden `file raro.gif` indica que el fichero `raro.gif` contiene una imagen de tamaño 12338 x 14385 cuando, en realidad, no contiene ninguna imagen. Explique de dónde sale que el tamaño de la imagen sea 12338 x 14385.

D. Baje de moodle a su directorio `lab4` los ficheros `glow.tiff` y `gmarbles.tiff`. Con la orden `file` obtenga los metadatos de cada uno de los ficheros y deduzca su significado. **Cópielo al formulario.**

Use la orden `hd` sobre cada uno de los ficheros (redirigiendo la salida hacia el comando `less`) para averiguar cuál es el número mágico de cada uno de ellos (8 primeros dígitos hexadecimales). **Copie al formulario los 8 dígitos hexadecimales y su correspondiente valor ASCII**. **Explique la relación entre el número mágico y los metadatos** (busque en Internet el significado de los 4 primeros caracteres del número mágico para los ficheros tiff).

E. Use la orden `hd` para averiguar cuál es el número mágico de un fichero ejecutable en Unix, contenido en los cuatro primeros caracteres. Por ejemplo aplíquelo al fichero que contiene la orden `cat`:

```
$ hd /bin/cat | less
```

Copie al formulario los 8 primeros dígitos hexadecimales y su correspondiente valor ASCII. Explique la relación entre el número mágico y los metadatos (busque en Internet el significado de los 3 últimos caracteres**).**

3. Manipulación de contenidos de ficheros mp3

El objetivo de esta parte de la práctica es aprender el uso de dos herramientas muy utilizadas para manipular los contenidos de ficheros de audio:

- `avconv`, conversor de audio y video, que permite manipular ficheros de audio y de video[5]
- `id3`, herramienta de edición de las etiquetas informativas (metadatos) que se incluyen en los ficheros mp3 para facilitar su catalogación, siguiendo el estándar ID3[6]

Para realizar esta parte de la práctica debe seguir los siguientes pasos:

A. Empiece por descargase un fichero mp3 disponible bajo licencia "Creative Commons". Para ello, abra el navegador y vaya a la página: http://freemusicarchive.org/curator/Creative_Commons y descargue, al directorio `lab4`, alguna de las canciones en formato mp3 (en el enlace subrayado de la canción, seleccione con el botón derecho del ratón "*guardar enlace como*").

[5]Más información en http://www.libav.org/
[6]Información adicional en http://www.id3.org

Cambie el nombre original por otro nombre más corto que no contenga espacios en blanco, manteniendo la extensión mp3 y utilizando una orden similar a la del siguiente ejemplo (¡comillas necesarias!):

```
$ mv "Beastie Boys - Now Get Busy.mp3" nom-fich
```

En el terminal de texto, averigüe el número de bytes que ocupa su fichero mp3 (por ejemplo con la orden `ls -l`). **Escriba en el formulario el nombre de su fichero mp3 y su número de bytes.**

B. La herramienta `avconv` admite muchas opciones y formatos. Para "cortar" el fichero mp3 que se ha bajado y generar otro fichero mp3 de menor duración, utilícela en la forma:

```
$ avconv -i nom-fich -ss inicio -t duración -acodec copy nom-fragm
```

en donde:

`-i` *nom-fich*	indica el fichero audio origen
`-ss` *inicio*	indica el instante en segundos en el que se quiere empezar a extraer
`-t` *duración*	la duración en segundos del fragmento a extraer (a partir del inicio)
`-acodec copy`	indica que el audio debe copiarse sin ninguna transformación
nom-fragm	es el nombre dado al fichero de audio mp3 resultante de la extracción

Ejecute la orden `avconv` poniendo los valores adecuados de las opciones para que "corte" el fichero *nom-fich* a partir de algún momento distinto del de inicio y obtenga un fragmento de duración 10 segundos, al que debe dar un nombre *nom-fragm* distinto del anterior pero con igual extensión mp3. **Copie al formulario de resultados la orden que ha dado para cortar el fichero.**

C. Ejecute ahora

```
$ avconv -i  fich
```

sobre ambos ficheros, el original *nom-fich* y el cortado *nom-fragm*, y podrá ver que en las dos líneas anteriores a la última donde se indica que debe proporcionar un fichero de salida, se muestran los metadatos, la duración, la tasa de bits (bitrate) y la frecuencia de muestreo del fichero de audio. **Copie al formulario de resultados estas dos líneas, para ambos ficheros.**

D. A continuación va a practicar con las utilidades `id3` e `id3v2` que facilitan la edición[7] de etiquetas en ficheros de audio siguiendo respectivamente las versiones 1 y 2 del estándar ID3 definido para incluir metadatos (etiquetas) en ficheros de audio. La versión 1 de ID3 establece las siguientes etiquetas, que se guardan en un bloque de 128 bytes al final del fichero mp3:

- Título (30 caracteres)
- Artista (30 caracteres)
- Álbum (30 caracteres)
- Año (4 caracteres)
- Comentario (30 caracteres)
- Género musical (1 carácter)
- Pista (valor entre 0 y 255)

[7]Existen también editores gráficos, como EasyTAG, que facilitan la edición de etiquetas tipo ID3 en los ficheros de audio. No se va a utilizar EasyTag en la práctica, sin embargo es recomendable que experimente con él.

La versión 2 de ID3 incluye las mismas etiquetas al final del fichero pero, además, añade otras etiquetas al comienzo.

Tecleando `man id3`, o simplemente `id3`, puede ver las opciones que se pueden usar para cambiar las diferentes etiquetas. (Lo mismo para `id3v2`)

Por ejemplo `id3 -a "Pepe Perez"` *`nom-fich`* cambiaría el nombre de la etiqueta *Artist* por "Pepe Perez". El valor de la etiqueta debe ir entre comillas siempre que haya espacios en blanco y no debe incluir caracteres no representables en ASCII (como vocales con tilde o ñ).

Ejecute `id3` con opciones `-l` o `-lR` para obtener los valores de las etiquetas ID3:

`$ id3 -l` *`nom-fich`* o `$ id3 -lR` *`nom-fich`*

Copie los valores de las etiquetas al formulario de resultados de la práctica.

Con la orden `hd` puede ver los contenidos del principio y del final del fichero y averiguar con qué versión de ID3 se incluyeron las etiquetas al crearse el fichero original. Dando la orden:

`$ hd` *`nom-fich`*`| head`

verá si el fichero comienza con 0xFFFB, en cuyo caso se tratará de la versión 1 de ID3v1, o comienza con "ID3" y, en tal caso, las etiquetas corresponderán al estándar ID3v2.

Ejecutando la orden:

`$ hd` *`nom-fich`*`| tail`

si el fichero tiene etiquetas ID3v1, verá que al final aparece "TAG", luego el título,…

E. Borre las etiquetas que hayan podido quedar en el fichero cortado, utilizando la herramienta `id3v2`, que elimina tanto las etiquetas ID3v1 como las ID3v2, con la orden:

`$ id3v2 -D` *`nom-fragm`*

Añada nuevas etiquetas ID3 personalizadas al fichero que contiene el fragmento. En concreto añada las etiquetas: *Artist* (poniendo las iniciales de su nombre), *Year* (el año actual) y *Comment* ("modificado por" seguido de sus iniciales). Una vez hechas las modificaciones, liste con la orden `id3` los valores resultantes y **cópielos al formulario de resultados de la práctica.**

4. Subida de ficheros de resultados a Moodle

Utilizando el navegador, suba dentro de la tarea "LAB4-G11. Entrega de ficheros de resultados." los dos ficheros de resultados de esta práctica:

- El fichero que ha ido rellenando durante la realización de la práctica, asignándole el nombre según el formato establecido: Grupo-Apellido1-Apellido2-L4.odt.
- El fichero que contiene el fragmento de audio con las nuevas etiquetas, asignándole el nombre según el formato establecido para las prácticas pero manteniendo, en este caso, la extensión mp3 (Grupo-Apellido1-Apellido2-L4.mp3).