

Convertir a UTF-8 viejas fuentes SGML

Autor José J. Grimaldos

La codificación de caracteres suele provocar verdaderos quebraderos de cabeza. Los sistemas configurados en nuestro entorno solían estar configurados para trabajar en ISO-8859-1, de manera que los ficheros fuente SGML editados en el pasado poseen esta codificación.

Todas las herramientas asociadas al trabajo con Dockbook que, a partir de estos ficheros fuente generan la documentación en diferentes salidas (HTML, PDF, TXT,...) están preparadas para esta codificación y, con ella, se obtienen los formatos finales, sin embargo, cuando estos documentos van a residir en gestores de contenidos o LMS, tipo Moodle, vuelven a aparecer los problemas con los caracteres, dado que estos sistemas ya están preparados para funcionar en UTF-8.

¿Qué hacer? Lo cierto es que existen varias alternativas para solventar esta circunstancia. Expondré, a continuación, la que me aportó Antonio Saorín, que funciona perfectamente y quede aquí como recordatorio.

En primer lugar, convertimos la codificación de los ficheros fuente con la orden (en una sola línea de la terminal):

```
iconv -f latin1 -t utf-8 archivo_ISO.sgml -o archivo_UTF8.sgml
```

Donde archivo_ISO.sgml denota el original codificado en ISO-8859-1 y archivo_UTF8.sgml será la salida del comando con el fichero convertido a la nueva codificación UTF-8

Ya está el trabajo hecho, ahora bastará generar los documentos en los distintos formatos ejecutando (en una sola línea de la terminal, cada una):

```
SP_ENCODING=utf-8 docbook2pdf -d ldp.dsl#print archivo_UTF8.sgml
```

```
SP_ENCODING=utf-8 docbook2html -d ldp.dsl#html archivo_UTF8.sgml
```

Es decir, anteponiendo SP_ENCODING=utf-8 a la instrucción habitual, para obtener las salidas PDF y HTML respectivamente. En este caso ldp.dsl es la hoja de estilos de The Linux Documentation Project con que se formateará el documento.