La estructura y manipulación de los ficheros PDF

Laurens Leurs

(Octubre de 1999)

En esta página se explica brevemente cómo se almacena la información en los ficheros de formato PDF:

Convenciones generales

esta información puede ser de alguna utilidad para quien quiera editar directamente ficheros PDF. Estos documentos son simplemente ficheros de texto ASCII de 7-bits. Se pueden abrir en cualquier editor de texto puro y duro (como el Bloc de notas de Windows, por ejemplo). siempre que el texto no se haya comprimido.

En un fichero PDF, cada línea puede contener hasta 255 carácteres. Cada una de estas líneas debe terminar con un carácter de retorno de carro (que debe ir o no seguido de un carácter de nueva línea dependiendo de la plataforma usada para crear el fichero PDF). En los datos un fichero PDF se establecen diferencias entre mayúsculas y minúsculas (es un formato *case-sensitive)*.

La estructura del fichero

El formato de ficheros PDF usa una estructura fija. Siempre contiene cuatro secciones:

1. Una cabecera *(header):* Contiene información sobre qué especificación del estándar PDF sigue el fichero. Esta información es algo parecido a "`%PDF-1.2` " (donde 1.2 puede ser 1.0 o 1.1 en el caso de las versiones más antiguas).
2. Un cuerpo *(body area):* Contiene la descripción de cada uno de los elementos usados en las páginas.
3. Una tabla de referencias cruzadas *(cross-reference table):* Contiene la información de qué elementos se usan en las páginas del fichero PDF.

4. Una coda *(trailer):* Que le dice al RIP dónde encontrar la tabla de referencias cruzadas y que termina siempre con un "`%%EOF`" (marca de final de fichero: *End Of File).* Si esta linea faltara, el fichero estaría incompleto y lo más probable es que un RIP no sea capaz de procesar el fichero. Esto no ocurre así con los ficheros PostScript, donde si falta la parte final del fichero (debido a un error de transmisión a una caída del sistema, etc...) aun es posible imprimir parte de la página (aunque sea de forma errónea). En un fichero PDF esto no ocurre. Se pierde todo.

Modificar los datos

Cada vez que se añaden nuevos datos a un fichero PDF (al editar por ejemplo un texto o al insertar nuevas páginas), se añaden sendas zonas nuevas de Cuerpo, Tabla de referencias y Coda. Si al guardar ese documento otra vez marcamos la opción "`optimizar`", Adobe Acrobat *limpiará* el fichero eliminando esas reduplicaciones de zonas y reorganizando el fichero.

Nota del traductor: Si quieres examinar y modificar la estructura de un PDF, existe un interesante programa llamado PDFCanOpener que te permitirá hacerlo.