

2. Documentos PDF

1. Introducción

PDF es un formato de archivo para representar documentos de una forma independiente de la aplicación software, hardware o sistema operativo que se haya usado para su creación, así como del dispositivo de salida por el que se muestra o imprime. Un documento PDF consiste en una colección de objetos que juntos, describen la apariencia de una o más páginas, posiblemente acompañados por elementos interactivos adicionales y datos de aplicación de alto nivel. Un archivo PDF contiene los objetos que componen un documento PDF, así como la información estructural asociada, todo ello representado como una única secuencia autocontenida de bytes.

Las páginas de un documento (y otros elementos visuales) pueden contener cualquier combinación de texto, gráficos e imágenes. La apariencia de una página se describe mediante un flujo de contenidos PDF, que contiene una secuencia de objetos gráficos para que sean representados en la pantalla.

Además de describir la apariencia estática de la página, un documento PDF puede contener elementos interactivos que son sólo posibles en una representación electrónica. PDF soporta *anotaciones* (annotations en lenguaje PDF) de muchos tipos, como notas de texto, enlaces de hipertexto, marcas, ficheros adjuntos, sonidos y películas. Un documento puede definir su propia interfaz de usuario; las entradas de ratón y teclado pueden activar ciertas acciones que están especificadas en los objetos PDF. Un documento también puede contener campos de un formulario interactivo para que sean rellenados por el usuario, y puede importar o exportar los valores de esos campos a otras aplicaciones (veremos más adelante que estas características no se pueden incluir en un documento para enviar al IEEE Xplore).

2. Componentes de un archivo PDF

Para conocer la estructura de un documento PDF nos remitiremos a la referencia de Adobe, en la que podemos comprobar como un documento PDF se divide en cuatro partes:

- Objetos: Un documento PDF es una estructura de datos compuesta por un pequeño conjunto de tipos básicos de objetos de datos.
- Estructura de Archivo: La estructura del archivo PDF determina como se almacenan los objetos en un archivo PDF, como se accede a ellos , y como se actualizan. Esta estructura es independiente de la semántica del objeto.
- Estructura de Documento: La estructura de un documento PDF especifica como se usan los tipos básicos de objetos para representar los componentes de un documento PDF, es decir, páginas, anotaciones, fuentes, etc.

- Flujos de contenido: Un flujo de contenido PDF contiene una secuencia de instrucciones que describen la apariencia de una página u otra entidad gráfica. Estas instrucciones, a pesar de que también se representan como objetos, son conceptualmente distintas de los objetos que componen un PDF, y por eso se describen aparte.

Dentro de los objetos, podemos encontrarnos con los siguientes tipos:

- Booleanos.
- Números Enteros y Reales.
- Cadenas de texto.
- Nombres.
- Arrays.
- Diccionarios.
- Streams.
- El Objeto Nulo.

Por su especial relevancia en el proyecto, describiremos los objetos de tipo diccionario. También haremos lo propio con los objetos indirectos, ya que nos ayudará a leer un archivo PDF por dentro, al abrirlo en modo texto.

3. Objetos de Diccionario.

Un Objeto Diccionario es una tabla asociativa que contiene pares de objetos, conocidos como Entradas de Diccionario. El primer elemento de cada entrada es la Clave, y el segundo elemento es el Valor. La clave debe ser un nombre, pero el valor puede ser cualquier clase de objeto, incluyendo otro diccionario. Una entrada de diccionario cuyo valor sea Null, es equivalente a una entrada ausente.

Dos entradas distintas de un diccionario no pueden tener la misma clave.

Un diccionario se escribe como una secuencia de pares Clave-Valor, delimitados por los símbolos << y >>. Veamos un ejemplo:

```
<< /Type /Example
    /SubType /DictionaryExample
    /version 0.01
    /IntegerItem 12
    /StringItem (una cadena de texto)
    .
    .
    .
>>
```

Los Objetos de tipo Diccionario son los pilares básicos de un documento PDF. Se usan

principalmente para agrupar y unir los atributos de un objeto complejo, como una fuente o una página de un documento, con cada entrada del diccionario especificando un valor para un atributo. Por convenio, la entrada *Type* de un diccionario identifica el tipo de objeto que describe el diccionario. En algunos casos, la entrada *Subtype* se usa para referirse a una subcategoría de un tipo más general. El Valor de las entradas Type o Subtype siempre es un nombre. Por ejemplo, un diccionario de Tipo Fuente, el valor de la entrada Type es Font, mientras que el subtipo puede ser Type1, TrueType, etc.

El valor de la entrada Type casi siempre se puede extraer del contexto. El operando de un operador Tf, por ejemplo, debe ser una fuente. Por lo tanto, la entrada Type sirve principalmente para documentación y corrección de errores. La entrada Type, por tanto, no es obligatoria a no ser que así se describa en el documento. Pero, de existir, ha de tener el valor correcto.

4. Objetos Indirectos.

Cualquier objeto de un archivo PDF puede ser etiquetado como un objeto Indirecto. Esto le proporciona al objeto un *identificador de objeto único* que pueden usar otros objetos para referirse a él, por ejemplo, como elemento de un Array o el valor de una entrada en un objeto Diccionario. Un identificador de objetos consta de dos partes:

- Un Numero de Objeto entero positivo. Los objetos indirectos a menudo se enumeran secuencialmente en un archivo PDF, pero no es necesario, pudiendo numerarse de forma arbitraria.
- Un número entero no negativo de generación. Se usa para las actualizaciones del documento.

Juntos, la combinación de un objeto número y un número de generación, identifican unívocamente un objeto indirecto.

La definición de un objeto indirecto en un archivo PDF consiste en su número de objeto, y un número de generación (por defecto, todos empiezan con número de generación 0), seguidos del valor del objeto, delimitado por las palabras clave **obj** y **endobj**. Por ejemplo, la definición:

```
12 0 obj  
    (Cadena)  
endobj
```

define un objeto indirecto con número de objeto 12, número de generación 0, y valor la cadena de texto “Cadena”.

De este modo, este objeto puede ser referido desde cualquier parte del archivo mediante

una *Referencia Indirecta* que consiste en el número de Objeto, el número de generación y la palabra clave **R**.

Veamos un ejemplo de uso de un objeto indirecto extraído de un archivo PDF real:

```
156 0 obj
<</PageMode/UseOutlines/Names 157 0 R/Outlines 183 0 R/Metadata 153 0
R/Pages 150 0 R/OpenAction[158 0 R/Fit]/Type/Catalog/PageLabels 148 0 R>>
endobj
```

5. Estructura de Archivo

La estructura de Archivo para un archivo PDF está compuesta, en principio, por los siguientes cuatro elementos:

- Una cabecera de una línea que identifica la versión de Especificación de PDF a la que se ajusta el archivo.
- Un cuerpo, que contiene los objetos, que conforman el documento que contiene el archivo.
- Una tabla de referencia cruzada que contiene información de los objetos indirectos en el archivo.
- Un *trailer* que da la localización de la tabla de referencia cruzada y de otros objetos especiales dentro del cuerpo del archivo.

Esta estructura inicial puede ser modificada por posteriores actualizaciones, que añaden elementos adicionales al final del archivo.

Cabecera
Cuerpo
Tabla de Referencia-Cruzada
Trailer

Las reglas hasta aquí descritas son suficientes para producir un archivo PDF bien formado. Aun así, pueden aplicarse reglas adicionales respecto a la organización de un archivo PDF

para permitir un acceso incremental eficiente a los componentes de un documento en un entorno de red. Esta forma de organización se llama Linearized, y será uno de los aspectos que comprobaremos de un archivo PDF.

1. Cabecera de Fichero

La primera línea de un archivo PDF es una cabecera que identifica la versión de la especificación PDF a la que se ajusta el archivo. Para un archivo que se ajusta a la especificación PDF 1.7, la cabecera debería ser:

%PDF-1.7

De todas formas, ya que cualquier archivo que se ajuste a una especificación anterior de PDF también se ajusta a la versión 1.7, una aplicación que procese PDF 1.7 puede aceptar archivos con cualquiera de las siguientes cabeceras:

%PDF-1.0 %PDF-1.1 %PDF-1.2 %PDF-1.3 %PDF-1.4 %PDF-1.5 %PDF-1.6

A partir de la especificación 1.4, la versión de la cabecera de archivo puede ser omitida en favor de la entrada *Version* en el diccionario de catálogo de documento. Esto permite a una aplicación creadora de PDF actualizar la versión usando una actualización incremental.

Bajo algunas condiciones, una aplicación de usuario podría ser capaz de procesar archivos PDF que se ajusten a versiones posteriores para las que se ha diseñado la aplicación. Las nuevas características de PDF se introducen por lo general de una forma en la que pueden omitirse de forma segura por una aplicación que no las entiende.

2. Cuerpo de Archivo

El cuerpo de un archivo PDF consiste en una secuencia de objetos indirectos que representan los contenidos de un documento. Los objetos de tipo básico descritos anteriormente, representan componentes del documento tales como fuentes, páginas, e imágenes.

3. Tabla de Referencia Cruzada (Cross-Reference)

La tabla de referencia cruzada contiene información que permite el acceso aleatorio a objetos indirectos dentro del archivo para evitar que sea necesario leer el archivo completo para localizar un objeto en particular.

La tabla contiene una entrada de línea-única para cada objeto indirecto, especificando la localización de ese objeto dentro del cuerpo del archivo. La tabla de referencia cruzada es la única parte de un archivo PDF que tiene un formato fijo, lo que permite que se pueda acceder de forma aleatoria a las entradas de la tabla.

Dicha tabla se compone de una o más *secciones de referencia cruzada*. Inicialmente, la tabla entera consiste en una única sección (o dos si el archivo es Linearized). Cada vez que el archivo se actualiza se añade una sección adicional. **(P93)**

4. Trailer de Archivo

El trailer de un archivo PDF permite a una aplicación que lea el archivo encontrar rápidamente la tabla de referencia cruzada y ciertos objetos especiales. Un archivo PDF debería empezar a leerse desde el final. La última línea del archivo contiene solamente el marcador de fin de fichero. Las dos líneas anteriores contienen la palabra clave startxref y el offset de bytes desde el comienzo del archivo hasta el comienzo de la palabra clave xref en la última sección de referencia cruzada.

La línea startxref está precedida por el “*trailer dictionary*”, que consiste en la palabra clave trailer seguida de pares clave-valor, delimitado el conjunto de pares por los símbolos << y >>.