La estructura de los documentos PDF

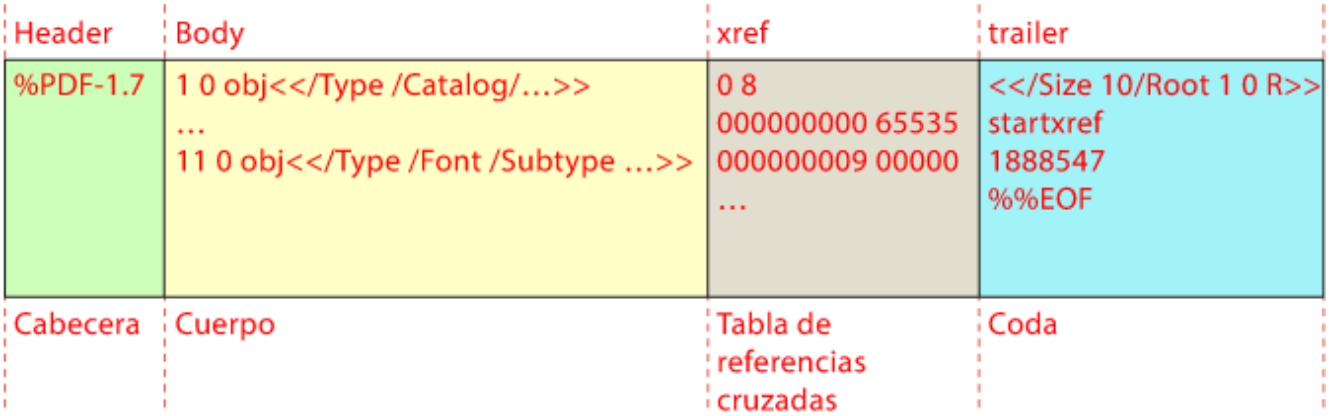
Gustavo Sánchez Muñoz

(Septiembre de 2022)

Un documento PDF es una estructura de código y datos que debe seguir unas reglas en su estructura y sintaxis para cumplir sus objetivos: Su finalidad es transmitir información manteniendo la integridad de la misma, especialmente en su presentación y formato. En origen no es un formato pensado para la edición.

Para ser independiente de la plataforma en la que se encuentre y mantener esa integridad, en principio y si no se ha comprimido o cifrado, un PDF es básicamente un archivo de texto ASCII, aunque incluya datos binarios (que son mucho más compactos).

Las partes de un PDF



Para ser válido, un documento PDF debe tener al menos y en este orden estas cuatro partes: (1) Cabecera *(header),* (2) cuerpo *(body),* (3) tabla de referencias cruzadas *(xref)* y (4) coda *(trailer).*

1. Cabecera *(header)*

La cabecera de un PDF lo identifica como documento PDF y declara a qué versión del formato se corresponde.

Consiste en una línea que comienza por "`%PDF-`" seguida de un número que identifica la versión del formato PDF (la menor es "`1.0`", la mayor es "`2.0`"); este

es un ejemplo de cabecera válida:

```
%PDF-1.7
```

Advertencia: A partir de la versión 1.4 del formato, Si en el diccionario llamado catálogo *(catalog)* se indica una versión distinta, se usará la del catálogo y no la de la cabecera. Sin entrar en detalles, éste es un procedimiento para garantizar la compatibilidad hacia atrás en algunos programas.

La presencia en la cabecera de una segunda línea que comienza con "%" y algunos caracteres binarios inmediatamente después tiene que ver con la presencia de datos binarios en el documento.

2. Cuerpo *(body)*

El cuerpo, la parte más importante de un PDF, contiene la descripción de cada uno de los elementos usados en las páginas. En un PDF, el cuerpo es lo que viene a continuación de la cabecera y termina con la tabla de referencias cruzadas. No tiene una marca de principio ni final ni un tamaño definido. Sus límites los marca la presencia de estas otras zonas.

La organización interna del cuerpo, que es absolutamente jerárquica pero aparentemente caótica, no se establece ni por el orden de aparición ni por los nombres de los objetos contenidos.

En el nivel superior de la jerarquía se haya un objeto raíz *(root),* que es el diccionario catálogo *(catalog).*

3. Tabla de referencias cruzadas *(xrefs table)*

Esta parte es una tabla con una entrada por cada uno de los objetos de programación usados en el documento, indicando su ubicación (por eso el cuerpo parece desorganizado).

La tabla de referencias cruzadas *(cross-reference table)* permite al programa que interpreta el PDF (para imprimirlo o mostrarlo en pantalla) acceder aleatoriamente en cualquier momento a cualquier elemento (matrices *(arrays),* valores booleanos (Boolean), valores numéricos, nombres (names), flujos (streams), cadenas de caracteres (strings) y diccionarios *(dictionaries)).*

Comienza identificándose con la línea "`xref`".

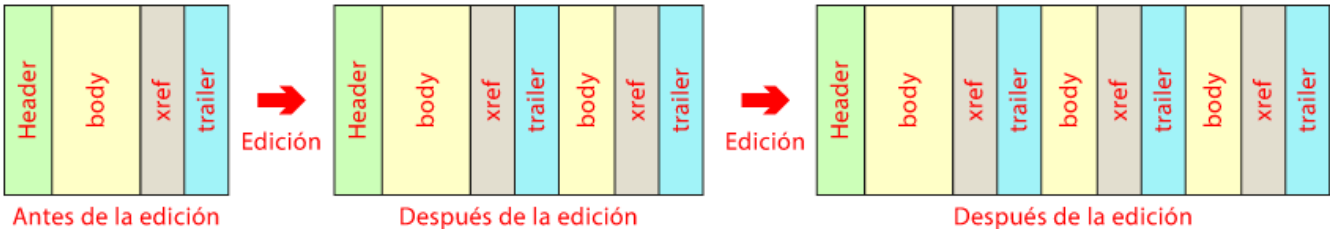
4. Coda *(Trailer)*

La coda de un PDF indica al programa que interpreta el documento dónde se encuentran algunos elementos esenciales para su lectura (como, por ejemplo, la tabla de referencias cruzadas y, usualmente, el objeto raíz *(root)* en la jerarquía del documento).

Comienza identificándose con la palabra línea "`trailer`" y termina con la línea "`%%EOF`" indicando "final del archivo" *(end of file).*

Incrementalidad y redundancia en la estructura de un PDF

Aunque el PDF no es un formato pensado para la edición de su contenido, sí se ha previsto la alteración de los documentos (para añadirle o quitarle páginas, para poner o quitar comentarios, etc.).



Eso quiere decir que es un formato que permite los cambios incrementales de la estructura. Cada vez que se alteran los datos y se guardan estos cambios, se crean nuevos cuerpo, coda y tabla de referencias cruzadas, que se colocan detrás de las anteriores (que permanecen).

Si estos cambios son numerosos, la estructura del documento puede volverse muy ineficiente. Los programas que permiten la edición de documentos PDF suelen tener la capacidad de optimizar la estructura de los archivos para eliminar esas redundancias y simplificar la estructura de nuevo.

Componentes básicos en un PDF

En el formato PDF, los datos se ordenan en estructuras básicas llamadas objetos *(objects),* como en muchas lenguajes de programación (entre ellos el PostScript),

Los principales de estos objetos son: Matrices *(arrays),* valores booleanos *(Boolean),* valores numéricos *(numbers),* nombres *(names),* cadenas de caracteres *(strings),* diccionarios *(dictionaries)* y flujos *(streams).* Estos dos tipos

son quizás los más importantes:

Los diccionarios *(dictionaries)*

En programación, un diccionario es una tabla de parejas en las que el primer término define una clave *(key)* y el segundo, el valor *(value)* asignado a esa clave. Un diccionario pueden estar formado por una sola pareja clave/valor o por una cantidad indeterminada de ellas. Posiblemente son el objeto más característico del formato PDF y muchos elementos usan este tipo de estructura.

Formalmente dentro de un PDF un diccionario se escribe con la forma:

```
111 222
obj<<
/clave1 /Valor1
/clave2 666
/clave3 [0 23 666 35 23 67 98]
>>
endobj
```

Donde la primera línea (con dos números "`111 222`") es el nombre del diccionario y "`obj<<`" y "`>>endobj`" definen el comienzo y final del diccionario. En su interior, cada línea forma una pareja clave/valor. La clave se define comenzando por la barra "`/`" y los valores se definen como cadenas de caracteres (que comienzan también con otra barra "`/`"), un número, una matriz (con sus valores entre corchetes), etc.

El valor de una clave puede ser, a su vez, otro diccionario (que se considera un subdiccionario). Buena parte de los datos y elementos de un PDF se organizan como diccionarios.

En un diccionario la clave "`/Type`" indica siempre en su valor el tipo de diccionario del que se trata; por ejemplo: "`<<…/Type/Pages…>>`" indica que ese diccionario contiene información sobre las páginas del documento. (También existen la clave "`/Subtype`"). No todos los diccionarios tienen porqué incluirlas.

Entre los numerosos diccionarios que pueden componer un PDF, destacan:

- **El catálogo *(catalog):*** Proporciona la referencia directa o indirecta a todos los objetos que forman el documento (aunque al final del documento, en la coda *(trailer)* puede haber algunos no referidos en el catálogo).

- **Las páginas *(pages):*** Las páginas son diccionarios *(page dictionaries)* que se organizan en un árbol de páginas *(page tree)* y donde se van definiendo o referenciando elementos menores como las imágenes, las fuentes, los textos, etc.

Los flujos *(streams)*

Son secuencias de bytes sin límite de tamaño prefijado capaces de contener grandes volúmenes de datos. Deben comenzar con un diccionario (donde se definen cosas como su longitud o el filtro de compresión usado), seguido de la palabra "`stream`", los bytes que componen el flujo y la palabra "`endstream`".

- **Los conjuntos de procedimientos *(ProcSets):*** En el interior de los documentos PDF, en los diccionarios de recursos de un flujo *(stream),* se pueden encontrar estructuras o elementos llamados "`ProcSet`". Éstos son operadores heredados por el formato PDF del lenguaje PostScript e indican conjuntos de procedimientos *(procedure sets),* operaciones que sólo deben usarse al imprimir en un dispositivo PostScript.

  Desde la versión 1.4 del formato PDF, estos procedimientos se consideran obsoletos y sólo se mantienen por motivos de compatibilidad con programas y dispositivos anteriores.