Los metadatos en un PDF

Gustavo Sánchez Muñoz

(Septiembre de 2022)

Los metadatos son datos que identifican a otros datos o conjuntos de datos; es decir: Son datos que sirven para especificar cuál es el propósito de una estructura de información.

En un símil sencillo: Si las estructuras de información fueran botes de comida y el documento las estanterías donde están esos frascos, los metadatos serían las etiquetas que indican qué hay en cada frasco, cuándo deben ser consumidos, qué calorías contienen, etc. No sólo facilitan el manejo de los elementos sin tener que analizarlos, sino que además proporcionan información extra que un análisis no necesariamente proporcionaría ("la mermelada favorita de mamá", por ejemplo).

En un documento PDF los metadatos pueden informar sobre el documento completo, estructuras y grupos de estructuras, donde cada uno puede llevar o no sus propios metadatos.

Advertencia: No hay que confundir la interactividad con los metadatos. Son ámbitos distintos. Sin embargo, en algunos programas se entienden también como metadatos algunas estructuras que ayudan a utilizar el documento proporcionando información sobre sus partes (anotaciones, miniaturas de página, información estructural, etc).

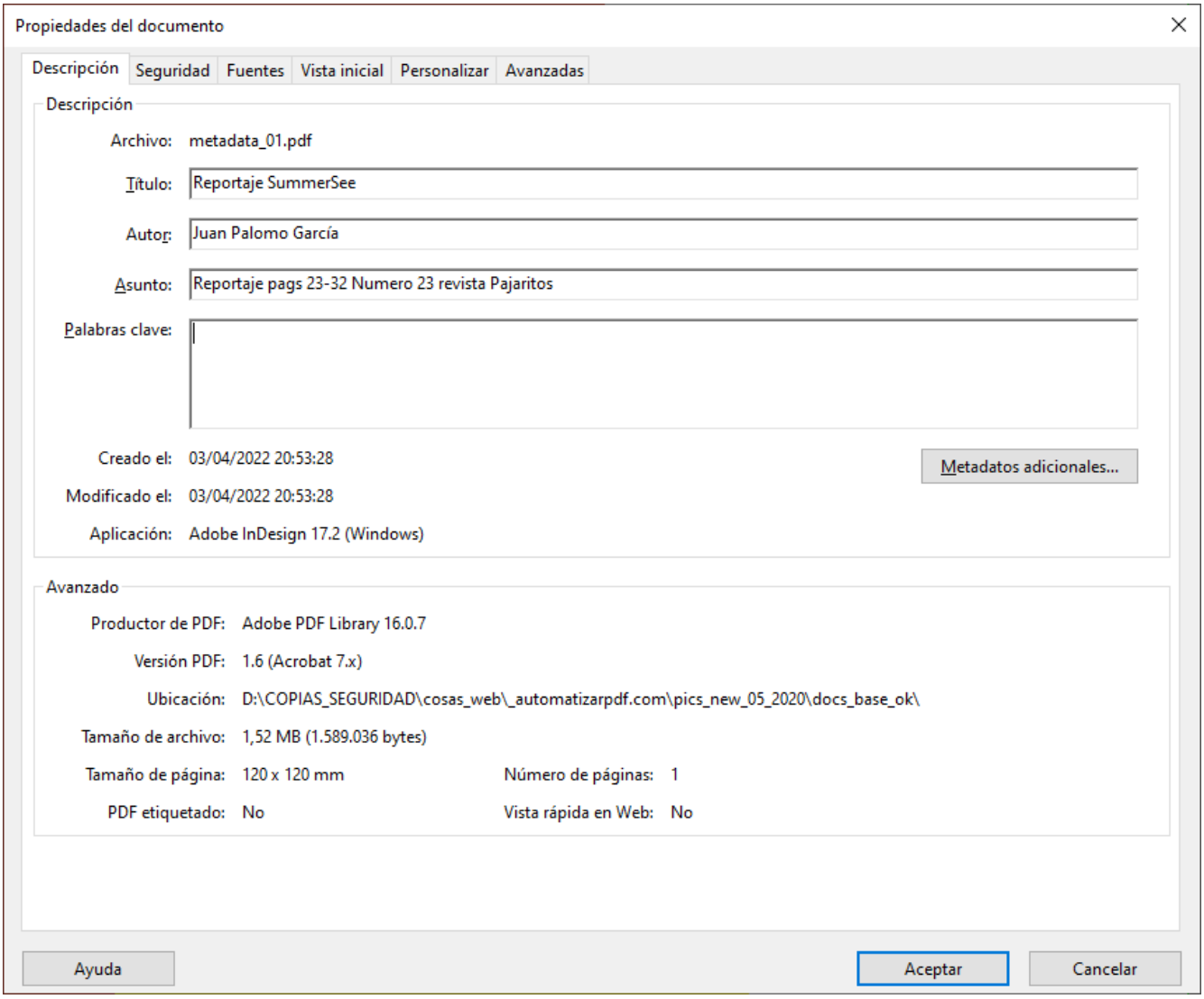
La incorporación de metadatos en las primeras versiones del formato PDF era bastante primaria, y se limitaba a cosas como el nombre del documento, quién lo

creo y cuándo. A partir de la versión 1.4 del formato, se añadió la posibilidad de etiquetar con metadatos los componentes individuales en el interior de los documentos (por ejemplo, las imágenes).

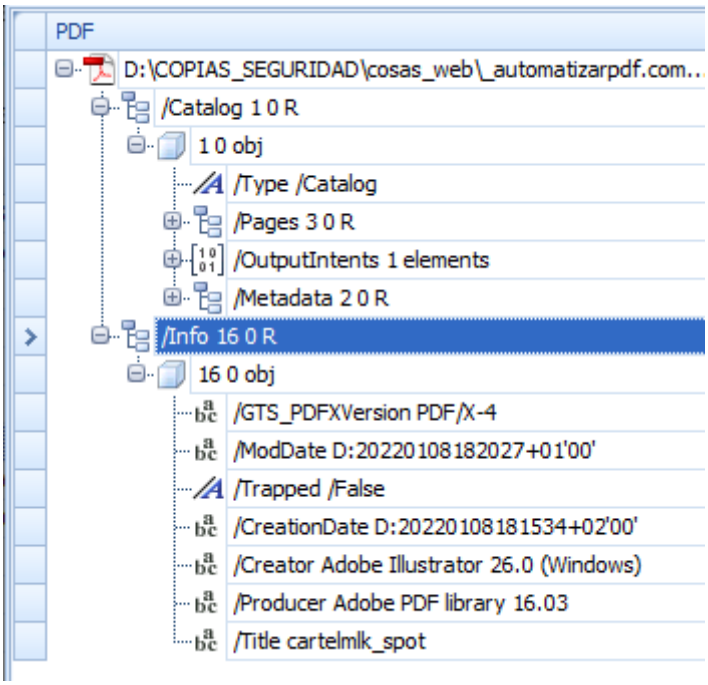
Formato de los metadatos en un PDF

En un PDF, los metadatos en sentido estricto se pueden añadir de dos maneras:

1. El diccionario de información del documento



Esta era la manera original de incorporar metadatos cuando se creó la primera versión del formato PDF. Sólo permite informar de datos generales sobre el documento. En Acrobat se puede acceder a ellos a través del menú "Propiedades".



En la coda *(trailer)* de un PDF se incluye una entrada llamada "`Info`", que es un diccionario de información del documento *(document information dictionary).* Allí, en forma de parejas "clave/valor" (como en todos los diccionarios) se puede incorporar la siguiente información opcional sobre el documento:

- **Título *(Title):*** El título del documento. En ausencia de otra instrucción, muchos programas usan el nombre del archivo a partir del que se creó el PDF.
- **Autor *(Author):*** La persona que creó el documento
- **Tema *(Subject):*** De qué trata el documento
- **Palabras clave (Keywords):** Palabras que identifican el contenido general.
- **Creador *(Creator):*** El programa con el que se creó inicialmente el documento (por ejemplo: "`Adobe InDesign 16.2 (Windows)`".
- **Productor *(Producer):*** Si el PDF se convirtió desde otro tipo de formato, el programa que hizo la conversión (por ejemplo: "`Adobe PDF Library 15.0`").
- **Fecha de creación *(Creation Date):*** Cuándo se creó el documento.

- **Fecha de modificación *(Modification date):*** Cuando se modificó por última vez el documento.

- **Reventado *(Trapped):*** Si el documento tiene reventado aplicado o no, o no se sabe ("`Desconocido`", que es el valor predeterminado).

El diccionario de información del documento permite la incorporación de otros datos siempre que se respete el formato clave/valor. Los programas que no tengan instrucciones sobre ellos los ignorarán. Algunos estándares añaden sus claves identificadoras aquí (aunque siempre van duplicadas y con mayor detalle en el correspondiente flujo de metadatos del catálogo del que se habla en el siguiente apartado).

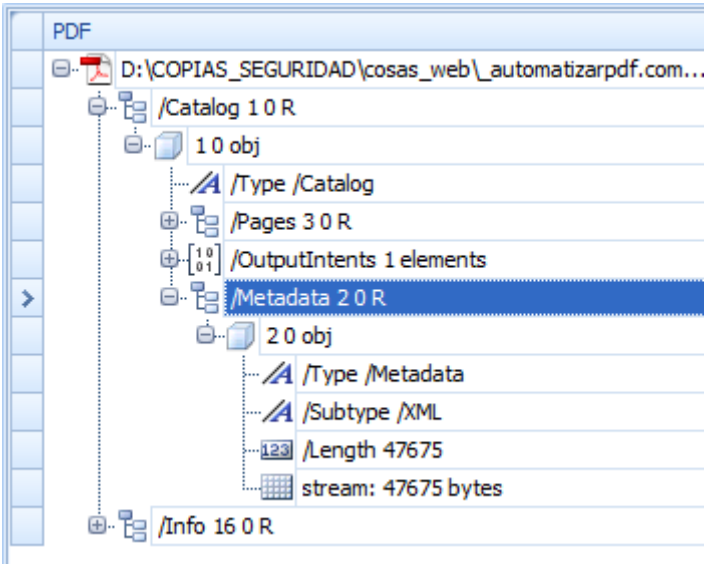
De esas claves, la principal es: "`GTS_PDFXVersion`", que indica que el documento es un PDF/X. La variante de PDF/X a la que pertenece se indica en el valor asociado (que sólo puede ser uno de los valores admitidos; por ejemplo: "`PDF/X4 `").

Advertencia: La existencia de este metadato o cualquier otro referido a un estándar no indica que el documento lo cumpla el estándar, sino sólo que sus creadores pretendieron que lo cumpliera.

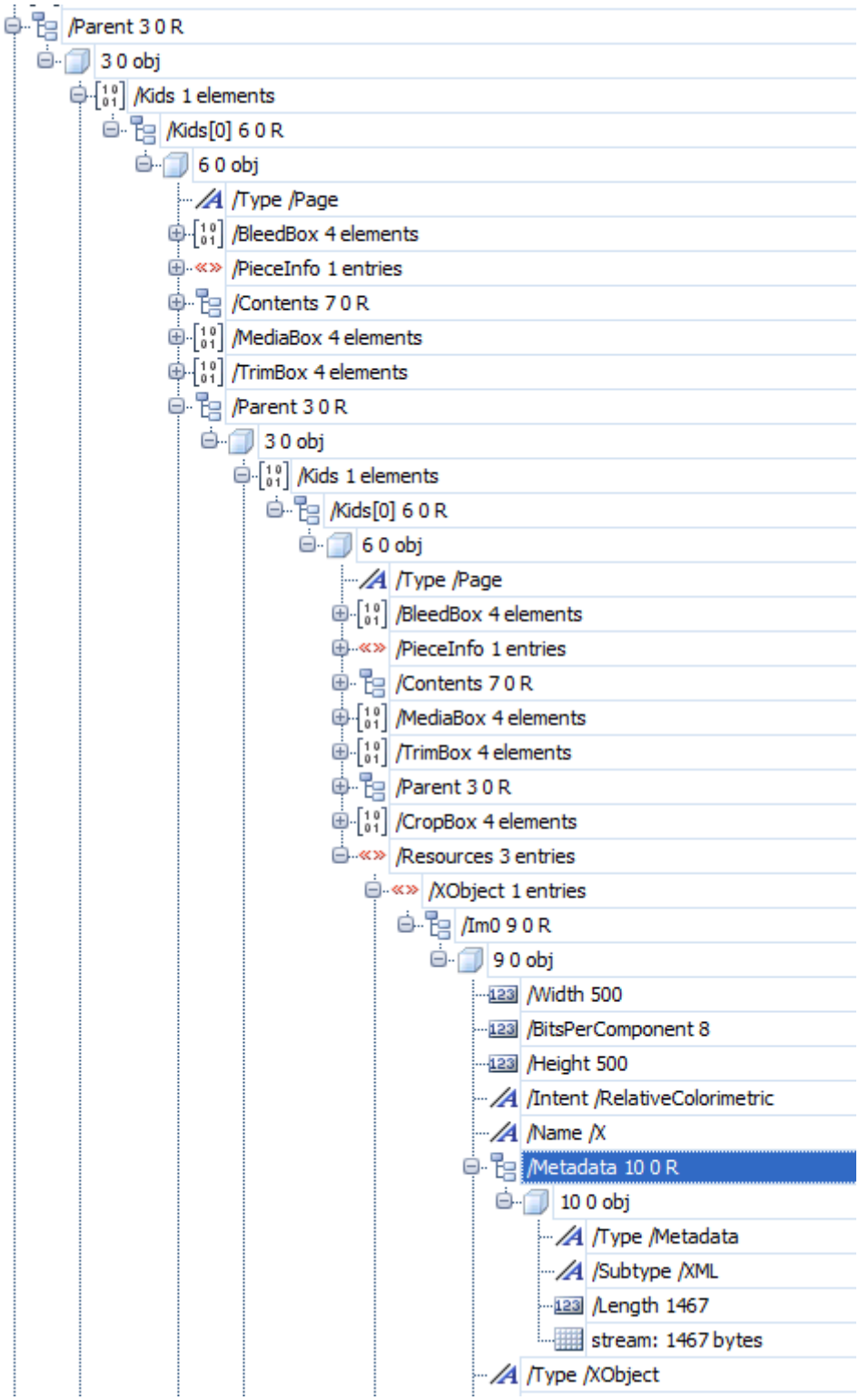
2. Los flujos de metadatos (metadata streams)

A partir de la versión 1.4 del formato PDF se añadió la posibilidad de añadir más metadatos en otras zonas del documento. Con este método, cualquier estructura de contenido de un PDF en forma de flujo *(stream)* o diccionario *(dictionary)* puede llevar metadatos adjuntos.

Se puede aplicar a elementos como páginas, formularios, fuentes o archivos incrustadas o perfiles de color, y, sobre todo, imágenes. Además, se puede aplicar al documento en sí mismo.



En este caso, cuando los metadatos se refieren al documento en general, deben ir como un flujo de metadatos en el diccionario llamado "Catálogo" *(catalog).* Es normal que esta información general duplique parte de la contenida en el diccionario de información del documento, que es en realidad una estructura obsoleta que se mantiene por motivos de compatibilidad.



Los metadatos de estructuras concretas, como por ejemplo imágenes, deben ir lo más cerca posible de ellas para evitar ambigüedades en su interpretación.

Los metadatos en estos flujos de metadatos deben presentarse siempre en una variante de XML llamada XMP, desarrollada por Adobe. Esto asegura que se identificarán e interpretarán correctamente.

Discrepancia de datos entre ambas zonas

Si hubiera diferencias entre los metadatos generales del diccionario de información del documento y los del flujo de metadatos del diccionario catálogo, prevalecen los datos que tengan la fecha de modificación más tardía.

PieceInfo

Los documentos PDF tienen la capacidad de incorporar datos privativos de otros programas. Eso ocurre, por ejemplo, cuando se le pide a Illustrator o Photoshop que guarde un PDF conservandolas capacidades de Illustrator. También ocurre con las versiones modernas de Adobe InDesign que permiten importar y exportar comentarios en los PDF.

Este tipo de datos se incorporan al PDF como metadatos mediante diccionarios llamados "`PieceInfo`", que se distribuyen por el documento.

Advertencia: Esta posibilidad de incorporar datos nativos de otras aplicaciones es la que dio origen a la idea de que PDF es un formato nativo de Illustrator y de que este programa era un editor de PDF. Ambas cosas no son ciertas.

Las especificaciones PDF/X y los metadatos

En sus primeras versiones, las especificaciones PDF/X sólo obligaban a la presencia de unos pocos metadatos generales: título de documento, fecha de creación y modificación. Además, se prohibía el valor "`desconocido`" en la clave "`reventado`" *(trapped))* y se obligaba a declarar el nivel del estándar PDF/X al que se atenía el documento.

A partir de la versión PDF/X-4 es obligatorio el uso de un flujo de metadatos general en formato XMP en el diccionario catálogo. Por el contrario, la presencia de esos datos en el diccionario de información del documento pasó a ser opcional (aunque si existen deben ser idénticos en ambas zonas).

Las especificaciones PDF/A y los metadatos

Es obligatorio que un PDF/A tenga un flujo de metadatos en formato XMP con la entrada "`metadata`" en el diccionario catálogo *(catalog)* del documento. El objetivo es garantizar que el documento contiene la información necesaria general sobre sí mismo.

Otras especificaciones de PDF y los metadatos

Los documentos PDF/E y PDF/UA tienen los mismos requisitos en sus metadatos que los PDF/A.