UNIVERSIDAD RAFAEL LANDÍVAR
CAMPUS DE QUETZALTENANGO
FACULTAD DE INGENIERÍA
MANEJO E IMPLEMENTACIÓN DE ARCHIVOS
SEGUNDO CICLO 2021
ING. DHABY XILOJ

PROYECTO SEGUNDA FASE
LECTOR DE METADATOS PDF

CARLOS RODOLFO SANTISTEBAN GRAMAJO #1559419
FERNANDO ENRIQUE QUIÑONEZ GARCIA. #1518619
LUIS MARIANO GUTIERREZ DIVAS. #1535719
MARCO JAVIER DE LEÓN VASQUEZ. #1521719

QUETZALTENANGO, 16 DE NOVIEMBRE 2021

Introducción

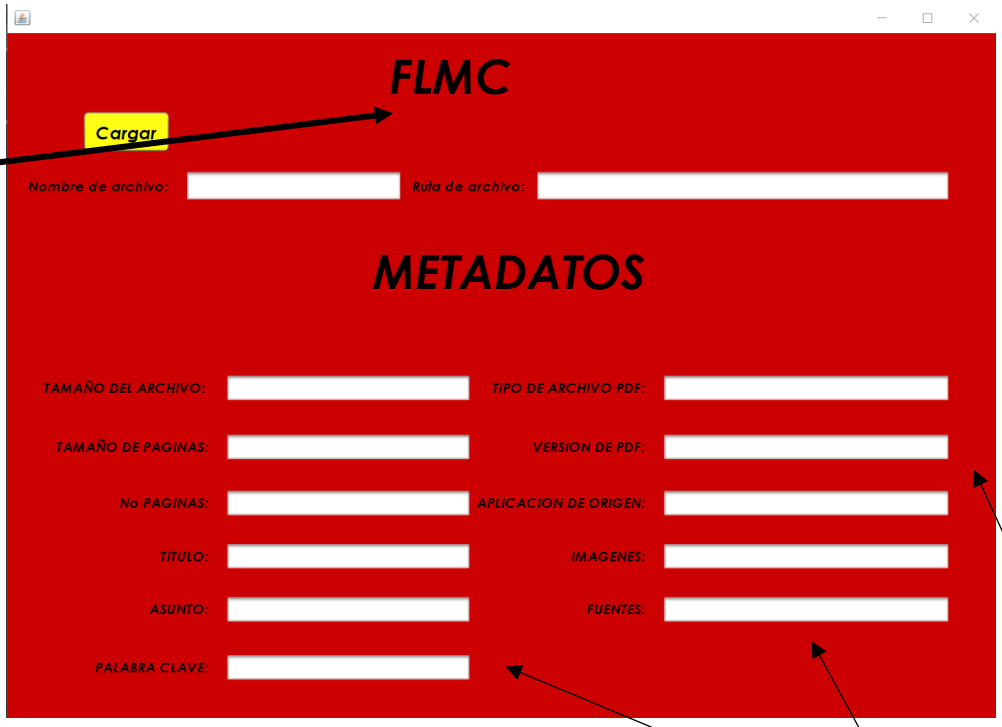
La aplicación fue creada a base de metadatos, que tuvimos que investigar con que estamos trabajando y encontramos que son aquellos datos utilizados para describir otros, por lo que son considerados los "datos sobre datos", ya que están en cualquier archivo que creemos y estos dan la información si lo hemos utilizado. Entonces la aplicación consiste en leer los metadatos que se encuentra dentro de los archivos PDF, los datos que leeremos son los siguientes: Tamaño del archivo, Tamaño de página, número de páginas, Título, Asunto, Palabras Clave, Tipo de archivo PDF, Versión de PDF, Aplicación con la que fue creada, Lista de imágenes que pueda tener el documento y Lista de fuentes que pueda incluir el documento. El programa nos dará la opción de elegir que archivo pdf queremos leer y este se abrirá leyendo todos los metadatos anteriormente dichos.

Como segunda parte estos metadatos ya han sido obtenidos de los PDF, y ya podemos guardarlos en un tipo de archivo llamado bin, que por medio del programa obligaremos a que los archivos se guarden en un orden en específico además de que estos datos se vayan juntando según el conjunto de metadatos que se esté abriendo y agregando a donde estén todos estos tipos de datos.

Manual de usuario

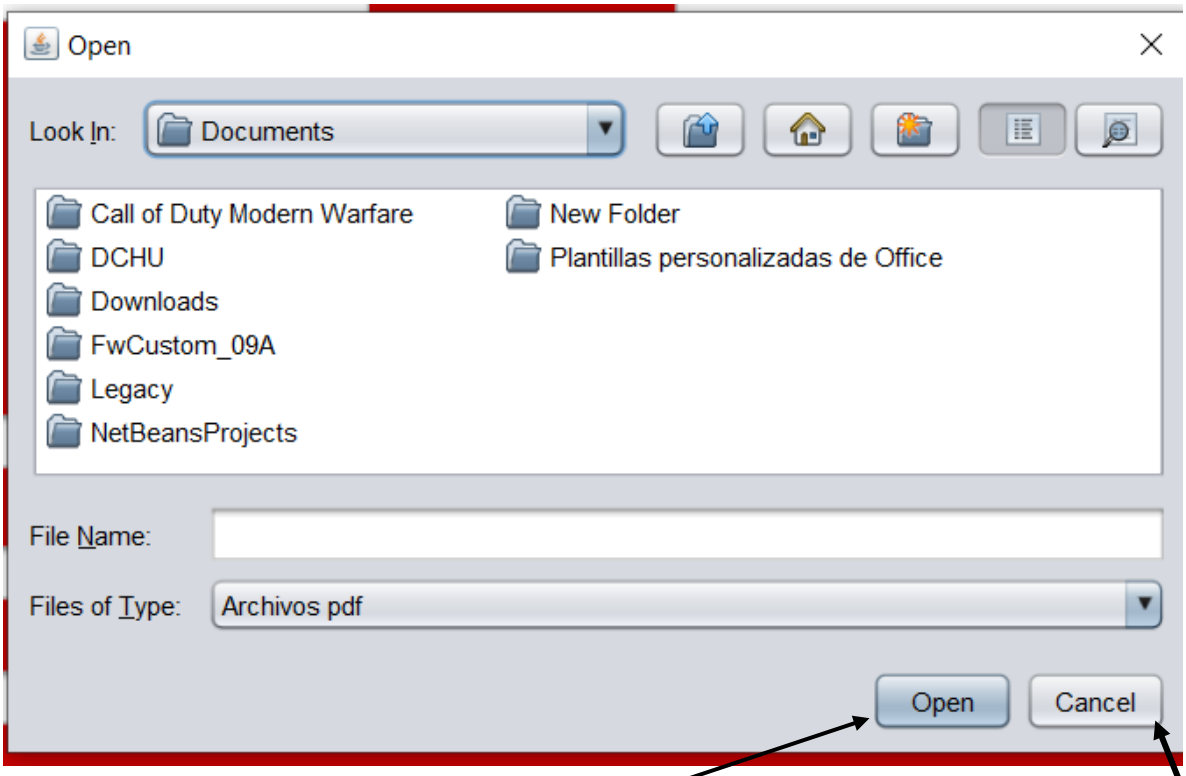
Al iniciar el programa se encontrará con esta pestaña en donde observa el logo del grupo, y también sirve para cargar los que son los documento PDF para mostrarlos.

Logo del equipo



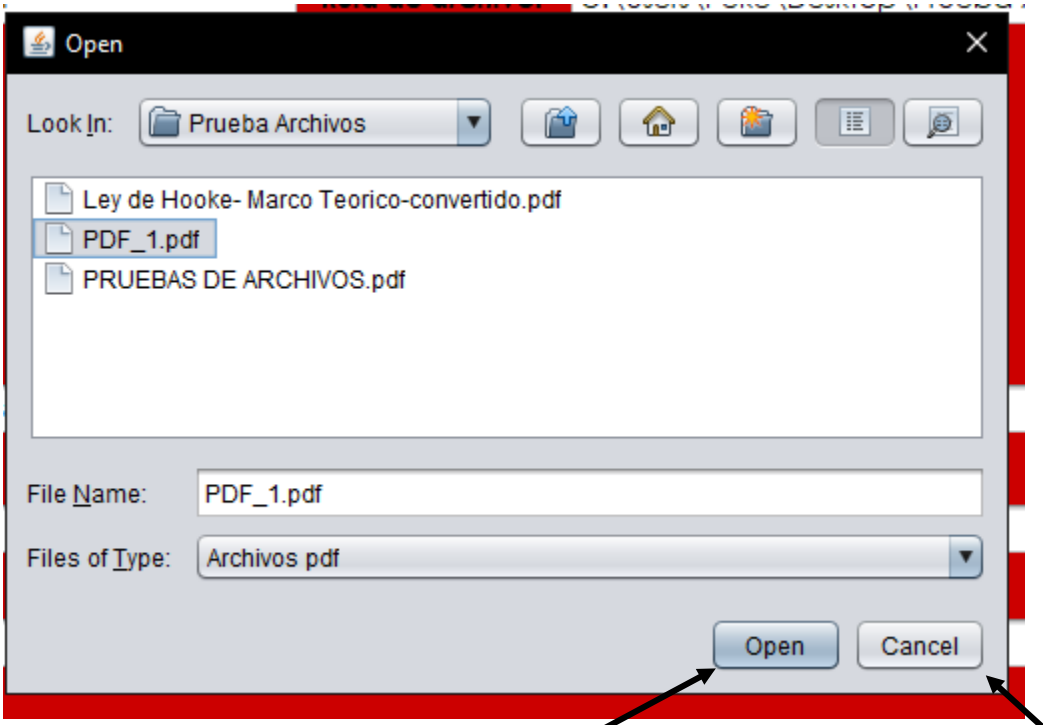
Espacios donde se mostrará la información

Con este botón “CARGAR”, abrirá una carpeta emergente en cual seleccionamos carpetas donde exista algún PDF.



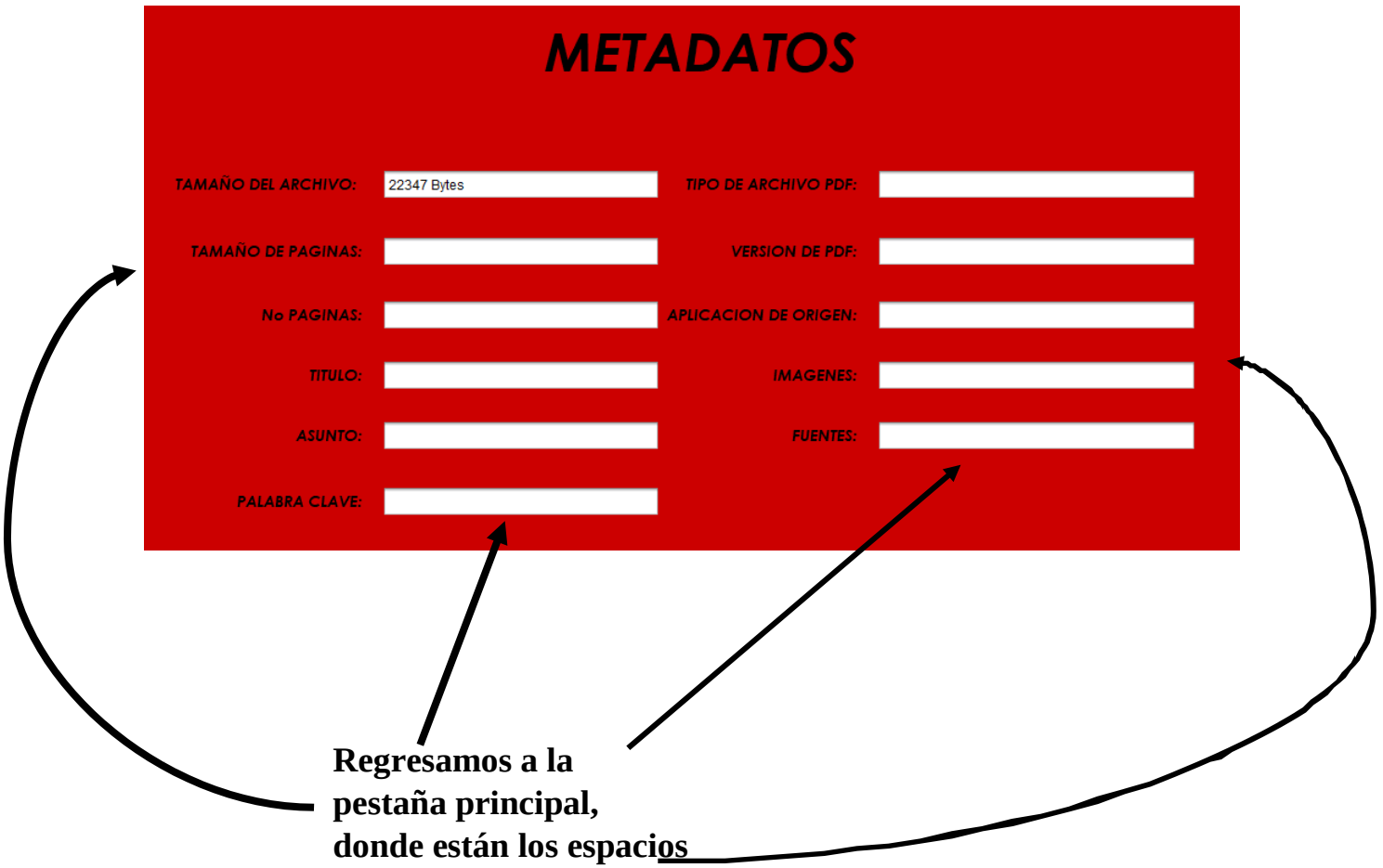
botón para abrir carpetas

botón en cual cancelamos



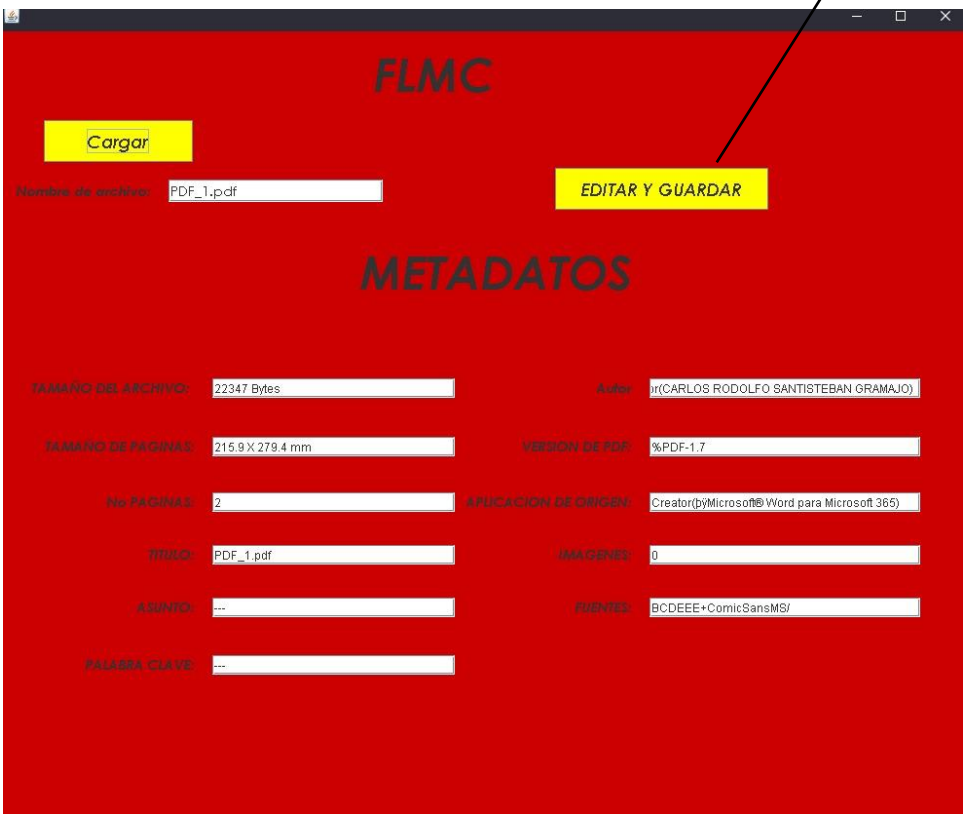
Con este botón confirmamos la carpeta a abrir

Cancelamos el archivo seleccionado.

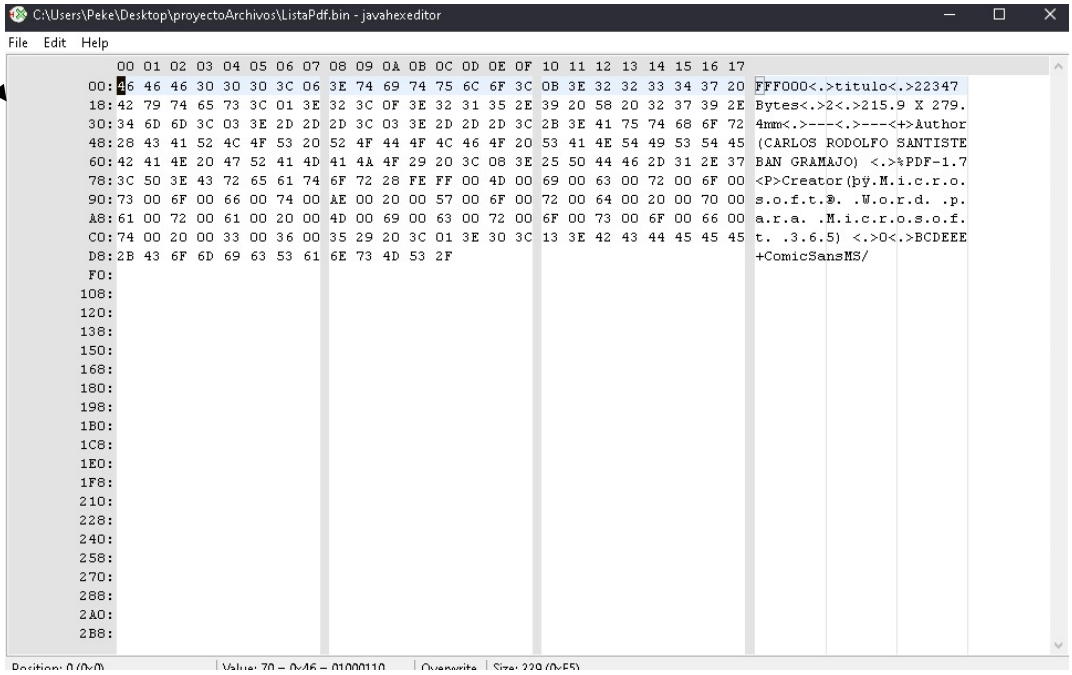


Regresamos a la pestaña principal, donde están los espacios para mostrar los datos del archivo PDF

Con este botón es de una nueva función el cual podremos editarlos PDF y guardarlos cuando terminemos



Aquí está el archivo de punto bin donde se irán almacenando los metadatos de los PDF.



Estructura de almacenamiento

El tamaño del archivo, tamaño de página, numero de página, título, asunto, palabras clave, tipo de archivo pdf, versión de pdf, aplicación con la que fue creado, lista de imágenes que puede contener, lista de fuentes que puede incluir el documento.

Short	Indica el tamaño del archivo que contiene la información	16 bits
Int	Indica el espacio entre cada tag. (versión, páginas, palabras clave, autor, título).	32 bits

Extracción:

Para poder extraer los metadatos de un pdf, primero debemos de conocer la estructura lógica del documento, luego debemos encontrar el diccionario del tráiler, que reside en el avance del archivo en el lugar del cuerpo principal del archivo, es una de las primeras cosas que se procesan cuando un programa desea leer un documento PDF. Contiene entradas que permiten leer la tabla de referencias cruzadas y, por lo tanto, los objetos del archivo. Una vez procesado el diccionario de avance, podemos continuar con la lectura del diccionario de información del documento y el catálogo de documentos.

En este documento se encuentran algunos metadatos como: el título, autor, fecha de creación y palabras clave.

El siguiente paso es entrar al catálogo de documentos, desde el cual se puede llegar a todos los objetos a través de referencias indirectas, como, por ejemplo: las páginas.

Sabiendo que nuestra tabla de referencias se encuentra al final del archivo (guiándose desde el editor hexadecimal), comenzaremos a leer los bits del pdf desde el final hacia el inicio, para poder ubicar la tabla y así poder encontrar los diferentes metadatos.

Guardado:

Para poder guardar los metadatos que ya han sido extraídos de los PDF. Estos pasan por un proceso el cual serán almacenados en un archivo de tipo bin. Lo cual estos datos son almacenados en u n orden en

específico. Y estos no solo contendrán los metadatos de un PDF si no que estarán agregados todos los metadatos de los PDF que se abran. Además que este solo se abrirá por una aplicación que lea la información de los diferentes tipos de archivos que existen y serán mostrados en hexadecimal.

Algoritmos utilizados:

Este programa fue desarrollado por la IDE NetBeans en el lenguaje de programación Java, para poder leer los datos de nuestro archivo, creamos una clase Leer, en la cual tenemos nuestros métodos para poder entrar a nuestra tabla de referencia y con esta poder ubicarnos en el último bit del archivo, y así obtener nuestra xref, y poder extraer los metadatos.

También contamos con JFrame visualizador que será la interfaz gráfica de nuestro programa, en esta podemos escoger el archivo PDF que queramos leer y también están los campos en donde se escribirán los metadatos extraídos del PDF.

Conclusiones

Al momento de realizar este proyecto aprendimos la importancia de conocer la estructura de los archivos, con la investigación que realizamos de cómo se guarda un archivo PDF, logramos trabajar de una manera más sencilla con la estructura del documento, y así leer los metadatos de cualquier tipo de PDF sin importar el programa de creación.

También es importante saber cómo está organizada la información en los archivos, y eso se logra gracias a una buena documentación ordenada. Lo cual nos resultó de mucha ayuda para guardarlo.

Se concede la realización de este proyecto para definir objetivos claros del aprendizaje que se lleva hasta el momento del curso llegando a abrir nuestra mente a muchos más conocimientos y experiencias en base a los archivos.

Se concluye que las funciones utilizadas en todo el proceso del proyecto que propone proporcionar una interfaz interactiva y fácil de usar para el usuario.

Enlace para visualizar el documento y la carpeta de nuestro proyecto en GITHUB:

- https://github.com/agentepeke/LectorMetadatosPDF/blob/main/proyectoArchivos/src/main/java/com/mycompany/proyectoarchivos/Leer.java