UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

FACULTAD DE CONTADURÍA Y ADMINISTRACIÓN

SISTEMA INTELIGENTE DE CAPTURA DE DATOS

DISEÑO DE UN SISTEMA O PROYECTO

QUE PARA OBTENER EL TÍTULO DE:

LICENCIADO EN INFORMÁTICA

PRESENTA:

CRISTIAN CARDOSO ARELLANO

ASESOR:

M. en I. LOURDES YOLANDA FLORES SALGADO

Cd. Mx. - 2019

UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

FACULTAD DE CONTADURÍA Y ADMINISTRACIÓN

SISTEMA INTELIGENTE DE CAPTURA DE DATOS

DISEÑO DE UN SISTEMA O PROYECTO

CRISTIAN CARDOSO ARELLANO

Cd. Mx. - 2019

Índice

Dedicatorias

A mi madre

Sra. Claudia Arellano

Gracias por los consejos y la experiencia que me has brindado, por estar ahí en los momentos más difíciles de mi vida.

Por tu amor y cariño inmenso hacia mí.

A mi padre

Sr. Marco Cardoso

Gracias papá por tu amistad y tus lecciones de vida.

Por ser mi brazo derecho, mi amigo y el mejor padre que jamás hubiese podido tener.

Tu comprensión y tu ayuda inmensa, si no fuera por ti esto no sería posible.

A mi abuela

Ma. Esther Escamilla

Gracias por ser mi madre, por tus lecciones de vida y sobre todo por siempre estar ahí en los momentos exactos cuando más te necesito.

A mi abuelo

José de Jesús Cardoso (1938 - 2013) ┼

Gracias abuelito por ser un consejero de vida,

todo lo que compartiste conmigo, los detalles, tus recuerdos,

tus consejos, tu alegría y sobre todo tu adversidad contra la vida.

Siempre estarás en mi corazón.

A mi compañero.

Adxel Ávila Mendoza

Por estar siempre ahí, por ser la guía y motivación de continuar.

A mi hermano

Josue Cardoso

Por tus consejos y sobre todo tu amistad.

El ánimo, las sonrisas y el apoyo que me brindas cada vez que lo necesito, por eso y por mucho más, gracias.

A mi mejor amigo

Ramón Alegría

Por tu amistad infinita, mi hermano.

Agradecimientos

A mi asesora

M. en I. Lourdes Yolanda Flores Salgado

Por ser mi instructora. Por sus enseñanzas y por ser mi guía, y sobretodo, por su apoyo incondicional.

A mi primo

Ing. Leopoldo Domínguez Cardoso

Por tu ayuda inmensa.

A mi amigo.
Lic. Tadeo Nava Martínez

Por su ayuda infinita en la redacción del presente escrito.

A la Facultad de Contaduría y Administración

Por brindarme los conocimientos para ser un profesional destacado en la industria de las tecnologías de la información.

Por ser el centro de mi formación académica y por permitirme conocer algunas personas muy valiosas para toda mi vida.

A la Universidad Nacional Autónoma de México

Por ser mi segunda casa, mi alma *mater*.

Introducción

"DS DATA DIGITAL S.A. DE C.V". en lo sucesivo "DS-DIGITAL", es una empresa que brinda soluciones digitales a organizaciones públicas y privadas de todo México. Dedicada principalmente a la digitalización de documentos, al almacenamiento físico de archivo muerto y al manejo de archivo digital.

Dentro de sus necesidades la empresa requiere de un sistema que automatice el proceso de captura de datos contenido en grandes archivos físicos y digitales. Posteriormente ordenarlos y renombrarlos utilizando los datos obtenidos de cada documento. La falta de este sistema conlleva altos costos y tiempos de operación.

El principal objetivo del presente trabajo es diseñar una solución factible llamada **SICD** (Sistema Inteligente de Captura de Datos). Dicho sistema implementa una solución al problema de una manera rápida y efectiva. Logra una reducción de costos y tiempos de ejecución hasta en un 70% de los que actualmente la empresa utiliza.

Implementar expresiones regulares[1] en el SICD es un objetivo específico y se propone su uso por su innegable poder y su enorme alcance para el procesamiento de texto, especialmente en documentos contables y legales.

El objetivo es desarrollar el SICD basado en un conjunto de herramientas de línea de comandos de *software* libre (*GNU General Public License)* y tecnologías *web.* Para así poder demostrar que se puede desarrollar software de calidad en México para brindar soluciones a bajo costo. La empresa decide no utilizar ninguna herramienta que actualmente se ofrece en el mercado; por la falta de personalización, módulos externos y precios elevados.

El SICD tendrá la capacidad para expandir los horizontes de la misma empresa. Creando un referente en lo sucesivo para enfrentar necesidades empresariales con

[1] Expresión regular: son patrones utilizados para encontrar una determinada combinación de caracteres dentro de una cadena de texto.

desarrollos eficientes y de bajo costo, sobre todo en el ámbito de la automatización de procedimientos relacionados a la digitalización y manejo de archivo digital.

En resumen los objetivos particulares del SICD son:

- Automatizar la separación de archivos.
- Automatizar la extracción de palabras claves por OCR.
- Automatizar la creación de una estructura de los archivos renombrados.

Capítulo I: DS DATA DIGITAL S.A. DE C.V.

Historia de la organización

DS-DIGITAL es una empresa mexicana de TI[2].

Nació en el año 2015 como persona física con actividades empresariales; con un proyecto de desarrollo de Software. En específico el desarrollo de la aplicación móvil de “Aranceles Edicomex” para plataformas con sistema operativo *Android*.

Tras este primer impulso la empresa continúa realizando desarrollos a la medida. En el año 2016 aparece una oportunidad para ofrecer un servicio de tratamiento y renombramiento de archivos digitales.

La empresa que requería el servicio, llamada Ancora Seguros, solicita el manejo automático de archivos digitales; específicamente un renombramiento de archivos basado en metadatos contenidos en los mismos. Se trataba esencialmente de pólizas de seguro de vida, hogar y de auto, que se contaban por varios miles. Por lo tanto la indexación automatizada era vital para el brindar el servicio.

Tras este servicio DS-DIGITAL se ve involucrado en la gestión de archivos digitales. Lo que permitiría a la empresa obtener experiencia en el terreno de los servicios de

[2] TI: Tecnologías de la información.

digitalización de documentos; rubro que se incluiría en los servicios ofrecidos a los clientes y que se ha ido extendiendo y perfeccionando en los últimos años.

Pronto DS-DIGITAL se encontró expandiendo sus servicios también en el almacenamiento de archivo, soluciones de seguridad en gestión y almacenamiento de información.

En el año 2018 DS-DIGITAL afirma en su página *web* contar con la experiencia de "50,000,000 documentos digitalizados, 120,000,000 documentos almacenados y 3,546,353 documentos renombrados" (DS-DIGITAL, 2018).

Objetivos

Entre los principales objetivos de DS-DIGITAL se incluyen:

Corto plazo (12 meses):

- Obtener la distribución de *CANON Mexicana S. de R.L. de C.V.*

Mediano plazo (3 años):

- Ser una de las cinco empresas líderes en el ramo de las soluciones digitales con especialidad en la digitalización de archivo y consulta de información.
- Contar con la certificación *ISO/IEC 27000*[3].

Largo plazo (5 años):

- Ser el líder en la digitalización de documentos en México.

Misión

DS-DIGITAL es una empresa nacida para brindar atención a pequeñas y medianas empresas y organizaciones en México que resuelve necesidades en la digitalización de documentos y automatización en el tratamiento de la información.

[3] ISO: Organización internacional de estandarización (*International Organization for Standardization*).
IEC: Comisión Electrónica Internacional (*International Electrotechnical Commission).*

Visión

Ser una de las empresas mexicanas más importantes en el rubro de la digitalización de documentos utilizando *software* hecho en México.

Valores

- La principal prioridad de DS-DIGITAL es la satisfacción de los clientes.
- Búsqueda de la calidad en el trabajo realizado.
- Honestidad.
- Eficacia, el compromiso y la responsabilidad.

Políticas

- Responsabilidad. La entrega en tiempo y forma de los servicios requeridos.
- Confidencialidad de la información proporcionada por los clientes.
- Garantizar siempre la disponibilidad de la información.

Clasificación de DS-DIGITAL

- ➢ DS-DIGITAL se clasifica por su tamaño como una PyME[4] pues cuenta con un capital humano de ocho personas de planta y un número variable de trabajadores eventuales según los proyectos.
- ➢ DS-DIGITAL se clasifica por el origen de su capital como una empresa privada.
- ➢ DS-DIGITAL se clasifica por sus actividades como una empresa del sector terciario ya que se dedica a proveer servicios.
- ➢ DS-DIGITAL cuenta con una página *web (https://ds-digital.com.mx)* desarrollada por la propia empresa.

[4] PYME: Pequeña y mediana empresa.

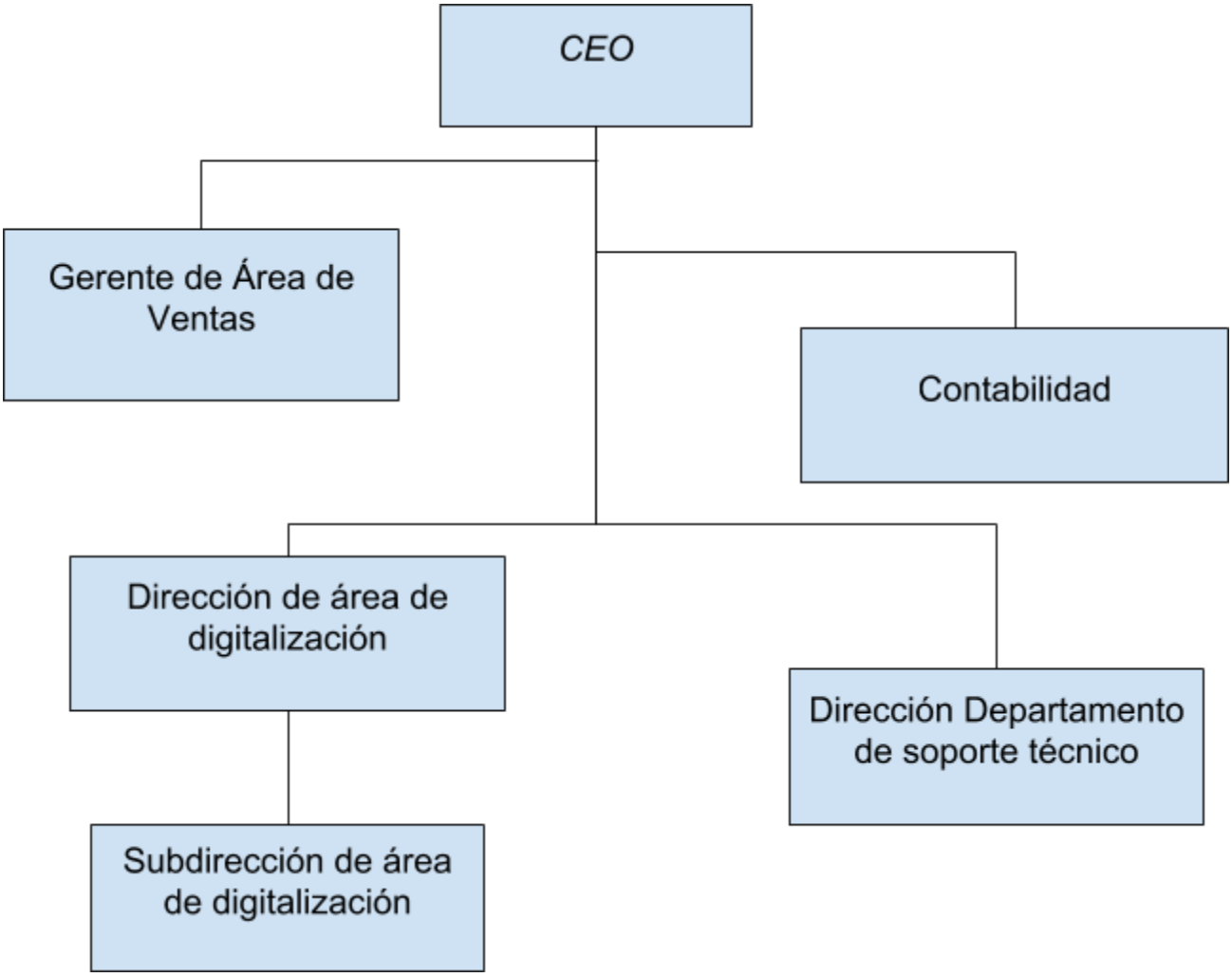

Fig. 1.1 Organigrama general DS-DIGITAL.

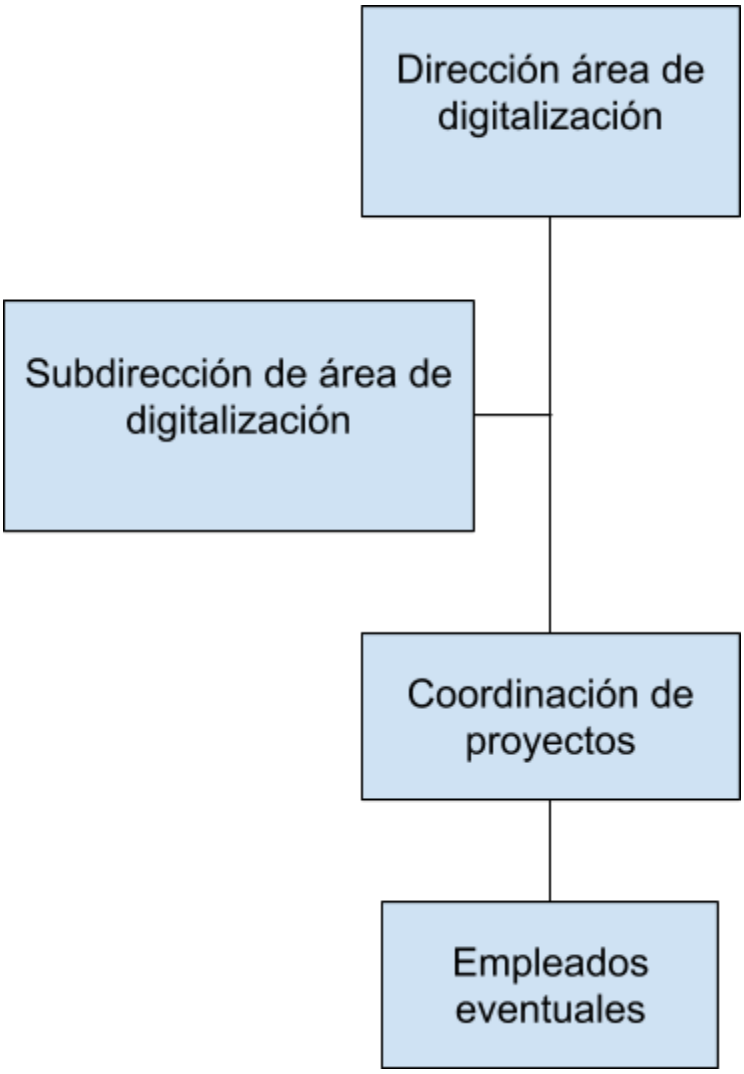

Fig. 1.2 Organigrama de área de digitalización DS-DIGITAL.

Capítulo II: Automatización de captura de datos en un proceso de indexación de documentos

Para entender lo que es un proceso de indexación y su relación con la captura de datos, y lo que conlleva su realización y automatización, es necesario entender el proceso de digitalización de documentos.

¿Qué es la digitalización de documentos?

La digitalización de documentos de texto o imágenes puede entenderse como un proceso de conversión desde un formato físico, casi siempre en papel, a un formato digital.

El objetivo principal de la digitalización de documentos es la posterior consulta de los documentos con diversos propósitos de una manera eficaz. Es importante tener en cuenta que para asegurar el posterior uso de los documentos digitalizados es necesario generar un proceso de información "**clasificada, estandarizada y con óptima calidad**" (*Archivo General de la Nación, 2015, p 7*) con el fin de requerir una sola digitalización.

Digitalización clasificada

Para clasificar la información en un proyecto de digitalización se requiere tener una organización de datos específicos extraídos y asociados a los documentos digitalizados. Estos datos se conocen como **metadatos** y la labor de organizarlos para ayudar en el acceso a la información se conoce como **indexación**.

No existe un estándar en cuanto a los datos susceptibles de extracción para una indexación correcta. Sin embargo los campos típicos a indexar son "número de factura, fecha, importes, nombres, direcciones, etc." (Grupo Telecon TBS, 2015).

La planeación de los metadatos permitirá hacer búsquedas eficientes de los documentos en el futuro y la posible integración de los documentos con otros sistemas importantes para las empresas como son los DMS[5], ERP[6] y los CRM[7].

Digitalización estandarizada

La estandarización en la digitalización de documentos se refiere a cumplir con los estándares ocupados para nombrar y organizar los documentos e imágenes digitalizados.

La manera de nombrar los archivos debe responder a la necesidad de que todos los documentos tengan manifestadas sus características más evidentes en el nombre de los archivos electrónicos, impidiendo duplicidades. Por lo cual es muy importante crear una clave o identificador construido a partir de metadatos estandarizados.

También es importante que la estructura de los archivos sea estandarizada para que la organización de la estructura digital no cambie entre un grupo de documentos y otro. Los criterios a definir para organizar una estructura de árbol son imprescindibles para garantizar que los documentos serán fácilmente recuperables en un futuro (*Wisconsin Historical Society*, 2018, p. 6).

Digitalización de óptima calidad

No existen estándares tecnológicos absolutos para el proceso de digitalización. Sin embargo existen instituciones públicas y privadas que tienen estándares recomendados. Por ejemplo el *Archivo Nacional de Québec* recomienda:

- La Norma *ISO 9660* para el registro y lectura de los datos en discos ópticos que tiene como fin de asegurar su migración a diferentes ambientes tecnológicos.

[5] DMS: Sistema manejador de documentos (Document Management System).
[6] ERP: Sistemas de recursos empresariales (Enterprise Resources Planning).
[7] CRM: Sistemas de gestión de relaciones con clientes (Customer Relationship Manager).

- El formato *TIFF*[8] para la toma de imágenes y los estándares *CCITT*[9] grupo 3 y grupo 4 para comprimir datos textuales.
- El formato *ASCII*[10] para la recogida de índices.
- La interfaz *SCSI-2*[11] entre el computador, el digitalizador y el grabador o quemador.
- Un monitor de visualización de alta resolución.

Por su parte la "Dirección de los Archivos de Francia" recomienda los siguientes formatos:

- *TXT* para textos.

- *XML* para textos estructurados.

- *R*ecomendación tamaño T4 para imagen fija en blanco y negro.

- *PNG* para imagen fija en color.

- Texto plano para base de datos.

(Muñoz, 2006, p.6)

El estándar tecnológico a utilizar puede ser el de alguna institución o bien uno propio creado específicamente para las necesidades de cada proyecto, siempre y cuando se cumplan dos condiciones:

- Que todo el proyecto de digitalización sea hecho con una calidad idéntica para cada tipo de documento.
- Que la calidad garantice que la información no se pierda y pueda ser utilizada en el futuro sin requerir una nueva digitalización (legibilidad).

[8] TIFF: (Tagged Image File Format).
[9] CCITT: Comité Consultivo Internacional Telegráfico y Telefónico (Consultative Committee for International Telegraphy and Telephony).
[10] ASCII: Código Estadounidense Estándar para el Intercambio de Información (American Standard Code for Information Interchange).
[11] SCSI-2: Interface de sistema pequeña de computadora (Small Computer System Interface).

El proceso de la digitalización

El proceso de digitalización consta de tres fases:

1. Preparación.
2. Digitalización o escaneo.
3. Indexación.

Preparación: En la fase inicial del proceso de digitalización. Se prepara el archivo documental para su manejo. No es buena idea digitalizar un archivo que no esté previamente organizado y clasificado. Es probable que se tenga que mover físicamente el archivo para llevarlo a un lugar donde se pueda trabajar con él, o en caso de trabajar en el lugar donde se almacena es necesario organizarlo en paquetes que estén distribuidos por fechas o algún otro identificador, y tienen que debidamente numerados. Finalmente se deben preparar los equipos de cómputo y los escáneres para su correcto uso. No es necesario preparar todo el archivo antes de empezar a digitalizar, pues basta con preparar una parte para iniciar con el escaneo y que lo restante se pueda ir preparando a la par de la digitalización (*ibid*, p.3).

Digitalización o escaneo: Es el procedimiento por el cual los documentos son escaneados, es decir, pasados por el escáner. Una de las principales cuestiones a considerar es el formato o tipo de documento que se quiere entregar, pueden ser documentos en formato *PDF* que almacenen grandes cantidades de documentos en un solo archivo, pueden ser imágenes en formato *TIFF* que pueden ser impresas fácilmente, o bien, formatos *JPG* o *PNG* si solo se van a solicitar en formato digital. Finalmente es muy importante considerar la estructura en cómo se almacena la información, ya que si el proyecto requiere estar estructurado, es en la fase del escaneo cuando la estructura se debe de generar (*EKCIT*, 2018).

Indexación: Finalmente debe hacerse la "captura" de los datos que permite identificar y acceder a los documentos correctamente. En este punto es donde se aplica la técnica *OCR*[12], que nos permite extraer texto de formatos de imágenes digitales.

[12] OCR:Reconocimiento Óptico de Caracteres.

La captura de los datos es de vital importancia, permite identificar los documentos en formato digital. La indexación puede ser tan sencilla como una tabla en una hoja de datos compuesto por de cinco a seis campos o tan compleja como un cúmulo de tablas en una base de datos. Sin una indexación adecuada se puede tener problemas en el producto final que se presentará al cliente si este requiere un acceso continuo y frecuente a la información (*ibíd, p12*).

Importancia de las técnicas OCR en la digitalización

Las técnicas *OCR* son métodos informáticos ya probados que permiten la extracción de caracteres por medio de análisis de imágenes. Estos caracteres son decodificados como palabras y mediante programación pueden ser determinados como metadatos de valor para la indexación en el proyecto.

Como se dijo antes, la indexación es una parte esencial para la clasificación de la información en todo proyecto de digitalización. Para ello es necesario implementar técnicas de *OCR* siempre que no se quiera incurrir en un problema de eficiencia.

Existen numerosas empresas que emplean capturistas para la labor de la extracción de datos. Sin embargo, en el "Diseño de procedimientos precisos de entrada de datos" de la Facultad de Contaduría y Administración de la *UAEM*[13] se refiere que con las técnicas OCR se puede ahorrar entre el 60% y el 90% del tiempo invertido en captura (*ITZELI*, 2015, p.49).

Es importante que en el proceso de preparación y captura se trabaje adecuadamente con miras a la implementación de las técnicas *OCR* para la indexación, ya que el nombre de los archivos, la estructura en que se almacenan los mismos y la manera en que están agrupados juegan un papel determinante en dicha implementación.

[13] UAEM: Universidad Autónoma del Estado de Morelos.

Limitaciones en el uso de OCR y alternativas

Las técnicas de análisis de imágenes tienen limitaciones. Para el caso del reconocimiento de caracteres, la limitación más importante es la casi nula posibilidad de captura de información contenida en letra manuscrita. Es muy importante saber que si hay documentos digitalizados que tengan letra manuscrita, como imágenes de documentos llenados a mano, conseguir efectuar la captura por *OCR* es muy improbable. Para este tipo de trabajos no hay más opción que habilitar un equipo de capturistas (*ONCE*, 2016).

Para documentos que nacieron digitales, o bien fueron llenados mediante máquinas de escribir, la captura por *OCR* está casi asegurada. Dentro de las diferentes herramientas informáticas para implementar *OCR* existen *plugins* que se adaptan a los diferentes tipos de letras de molde que existen.

En caso de que el proyecto sea incompatible con las técnicas de *OCR*, una alternativa para mejorar la eficiencia de la captura de datos consiste en crear un *software* que agilice el proceso de captura. Si bien, el resultado no es el mismo en términos de eficiencia, si se puede mejorar el rendimiento de los capturistas.

Uso de códigos en los procesos de digitalización

Para la indexación y la construcción de la estructura es habitual el uso de códigos.

Al utilizar códigos para la indexación y la creación de estructuras de almacenamiento de documentos se obtienen grandes ventajas en el manejo de archivos.

Otorgan:

1. **Unicidad**: Es factible asegurar que todos los documentos tengan un código único, por lo menos en cuanto a sus documentos “hermanos”, es decir, aquellos que se encuentran en su mismo directorio. Lo mismo aplica para los directorios. De esa

manera, los archivos gozarán de un identificador único combinando su nombre y el de su directorio padre (*ibid*, p.34).

2. **Clasificabilidad:** El código debe contener información sobre la clasificación del documento. Estas clasificaciones pueden ayudar a recuperar el documento de manera sencilla y ayudan a implementar herramientas de *software* que puedan determinar un tipo de documento específico para ejecutarse (*ibid*, p.14).
3. **Confidencialidad**: Los códigos ayudan a mantener información disponible para los sistemas y usuarios de una organización a la vez que están ocultos o semiocultos para agentes externos. Si bien, la mayoría de los códigos pueden inferirse, sólo aquellos que tengan las claves de decodificación podrán hacer un uso eficiente del código para identificar los datos de cada documento (*ibid*, p.19).
4. **Dirección de función**: Los códigos también ayudan a identificar documentos que se encuentran dentro de un proceso. Parte del código puede mostrar el estado del documento, así como en qué procedimiento se encuentra en determinado momento. Con ello podemos reconocer archivos a los que se les va a realizar alguna acción (*ibid*, p.23).

Los códigos más utilizados son los llamados “Código de Secuencia Simple”, que son una secuencia de caracteres, subpalabras yuxtapuestos con significados codificados. En este tipo de códigos lo más importante es saber el orden de los caracteres o subpalabras para identificar la información que el código está mostrando. Cada carácter o subpalabra tiene su propia codificación, por lo que se necesitan varios diccionarios de claves para revelar la información completa del código. Si no se desea ocultar información, también es posible utilizar palabras completas con significados obvios.

Mario Rizo Rivas en su artículo para la revista *Forbes* titulado “Los eficaces son los nuevos sabios” explica la importancia de la eficacia y los estándares para el mundo empresarial y organizacional moderno como una clave en el éxito profesional. Es por ello que para el proyecto de digitalización el éxito ésta en cumplir con el cometido de digitalizar el archivo en cuestión sin descuidar ninguna de las tres características

anteriormente descritas (digitalización clasificada, estandarizada y de óptima calidad), todo de una manera eficiente.

Además de la eficiencia, existen otros elementos a tomar en cuenta para que un cliente de digitalización resulte satisfecho en sus necesidades. Por ello, antes de digitalizar un archivo se debe plantear al cliente las siguientes preguntas:

¿La información digitalizada será consultada todos los días?

Esta pregunta es importante para conocer el nivel de organización que requerirá la información en digital y el o los métodos para acceder a ella.

¿La información necesitará protección especial?

Si bien toda la información debe encontrarse protegida de agentes externos que puedan comprometer la seguridad de la institución a la que pertenece la información existen restricciones internas, ya que puede existir información a la que no todos los usuarios deben poder acceder. Esto también afecta a la manera de indexar y organizar la información.

¿Se destruirá la versión original en papel después de la digitalización?

En el caso de que la versión en papel sea destruida después de la digitalización se podrá crear copia alguna en versión digital, por lo que un *backup*[14] de la información es imperativo (Bastin, Hurtaud, & Senequier, 2014, p.81).

Los diferentes tipos de servicio de digitalización pueden generar varios tipos de negocio, como son:

- Digitalización inmediata

Consiste en un servicio pequeño, en el que la organización es mínima y se entrega al cliente la versión digital de su información por algún medio de almacenamiento de uso común, por ejemplo en una memoria flash *USB*.

[14] Backup: Copia de información digital resguardada en un dispositivo de almacenamiento.

Requerimientos en Hardware:

- Equipo(s) de cómputo.
- Escáner.
- Unidad de almacenamiento.

Requerimientos en Software:

- *Software* de digitalización.
- Puede requerir una funcionalidad *OCR.*

Otros Requerimientos:

- Ninguno.

- Bóveda de datos

Es un servicio en el que la información se encuentra resguardada y almacenada tanto física como virtualmente. Se entiende que la información no va a ser requerida de manera continua ni frecuente. La información será tratada como se trataría un archivo importante almacenada en una bóveda de seguridad. Mientras que no es importante la creación de un sistema de interacción de la información, la seguridad física de los dispositivos de almacenamiento es vital (*ibíd*).

Requerimientos en *Hardware*:

- Unidades de almacenamiento con alta seguridad.
- Servidor o clúster[15] de servidores.
- Equipos de cómputo y escáneres.

[15] Cluster: cúmulo, granja o cluster de computadoras, sistema de procesamiento paralelo o distribuido que consta de un conjunto de computadoras independientes, interconectadas entre sí.

Requerimientos en *Software*:

- Sistemas avanzados de seguridad.
- *Software* para digitalizar.
- Puede requerir una aplicación *OCR* para extraer datos.

Otros requerimientos:

- Instalaciones con equipo de seguridad.

- Sistema manejador de documentos (*DMS*)

El caso más común de negocio es el *DMS*. Consiste en un sistema que permita a la vez almacenar y gestionar el acceso frecuente a la información por parte de un grupo de usuarios que pueden, o no, tener los mismos privilegios de acceso. Para este caso deben preverse fallos humanos en el manejo de la información, y en algunos casos, actualizaciones periódicas de la misma (*ibíd*).

Requerimientos en Hardware:

- Equipo de cómputo y escáneres.
- Servidor local o *web*. Puede, o no, estar instalado en las instalaciones del cliente.
- Unidades de almacenamiento de información.

Requerimientos en *Software*:

- Software de digitalización.
- Sistema de gestión de información.
- Sistema de obtención de datos mediante *OCR* .

Otros Requerimientos:

- Equipo de desarrollo para la implementación, implantación y adecuación de los sistemas requeridos.

- Archivo electrónico

Una combinación entre la bóveda de datos y el *DMS*. Trata sobre el almacenamiento de información legal o de negocio de alguna institución o empresa. No suele requerir actualizaciones ni un sistema de acceso elaborado, simplemente se debe tomar en cuenta que en algún momento se podrá requerir la información tal cual se guardó. Debe tener un protocolo estricto en caso de que se requiera eliminar la información. Existen documentos con llaves electrónicas que deben ser almacenadas en la indexación. En cuanto a seguridad, lo más importante es que los documentos no se pierdan con el tiempo (*ibíd*).

Requerimientos de *Hardware*:

- Equipos de cómputo y escáneres.
- Unidades de almacenamiento.
- Servidor o clúster de servidores *web.*

Requerimientos de *Software*:

- Software de digitalización.
- Sistema de obtención de datos mediante *OCR.*
- Sistema para el acceso a la información.

Otros requerimientos:

- Desarrollador para adaptar el sistema *OCR.*
- Equipo de desarrollo para implementar un *software* de acceso a la información seguro, con los protocolos para documentos legales.

Es importante hacer notar que en todos los modelos de negocio se puede requerir *OCR*. En el primero podría ser un desarrollo, en especial si los documentos cumplen con el mismo formato. En los posteriores modelos la complejidad de la herramienta se dispara y resulta de gran dificultad hacer uso de una herramienta genérica, ya que los datos que deben ser capturados pueden estar en cualquier parte del documento.

Planteamiento del problema: Necesidad de una aplicación OCR en DS-DIGITAL

En DS-DIGITAL surgió la necesidad de utilizar *OCR* con la aparición de proyectos de digitalización e indexación de miles de documentos. Para el presente documento se expondrá el proyecto específico con una compañía aseguradora “A” la cual **entrega** un promedio de 10,000 documentos mensuales, entre documentos físicos para ser digitalizados y documentos digitales que requieren de una indexación adecuada utilizando datos que están contenidos como información dentro de los archivos digitales.

Los archivos en físico constan de paquetes de documentos, en su mayoría pólizas de seguros.

Los archivos digitales constan de grandes archivos en formato *PDF*, con una cantidad de páginas que oscila entre 600 y 4,500 cada uno.

La solución esperada por el cliente es el alojamiento de los archivos digitales separados por pólizas en la nube de DS-DIGITAL. Indexados de tal manera que se permita su recuperación. Tanto los archivos digitales como los físicos se almacenarán juntos.

Para nombrar y organizar los archivos se utilizan datos contenidos como información en los mismos documentos. Los datos para la indexación y la estructura son variables y se definen por el cliente al momento de requerir un nuevo servicio.

Ejemplo de una entrega

Una póliza como la que se muestra en el “Anexo A” forma parte de un paquete de las mismas en la que el documento consta de más de 4,000 páginas que requieren ser separadas por pólizas y con un o varios identificadores únicos.

Cada póliza debería guardarse en un directorio nombrado dentro del cual se generará el *PDF* de la póliza nombrado con el siguiente código de secuencia simple:

[número de cliente][número de póliza][número de certificado].pdf

donde el número de cliente y el número de certificado son junto con el número de póliza los datos que vienen contenidos como información en los documentos. Como se puede observar, se trata de un código hecho mediante la yuxtaposición de palabras.

Al momento de la realización del presente documento los directorios de cada póliza se encuentran dentro de otro directorio general nombrado en referencia a la entrega de cada pedido sin requerir una estructura compleja, aunque en un futuro se planea implementar un *DMS* que funcione como *un cliente*[16] para dar acceso a los archivos a ciertos usuarios de la aseguradora A, lo que implicaría la utilización de una estructura más compleja.

Acotación del problema

El cliente requiere una solución con el servicio en una forma fácil y económica, teniendo la seguridad de que va a estar propiamente dividida, clasificada y organizada manera útil a sus propios intereses. El uso del código con el que se renombran los archivos tiene que ver con el *ERP* que maneja la aseguradora. Las fases en el manejo posterior de la información no son de incumbencia de DS-DIGITAL y el trabajo que se realiza solo se limita a cumplir con los requisitos del cliente.

El procedimiento de recuperación de la información es bastante sencillo, simplemente se le asigna una cuenta de usuario al cliente dentro de la nube de los servidores de DS-DIGITAL, con permisos para acceder a los documentos que le corresponden. La nube que está habilitada es el aplicativo *ownCloud*[17]. El servicio de la nube como tal es parte del

[16] Cliente: es una aplicación especializada en un tipo determinado de conexión que soporta un protocolo específico.
[17] ownCloud: aplicativo web (nube de aplicación) donde se comparte y sincroniza todo tipo de archivos con clientes para escritorio, web y dispositivos móviles.

problema que se encuentra resuelto y no formará parte de los capítulos posteriores del presente documento.

El problema de la logística para el trabajo de los archivos físicos que abarca la organización del archivo en las instalaciones del cliente y el transporte para recoger archivo y para la entrega del mismo está debidamente resuelto. El área de digitalización se encarga de la logística de transporte, ya que la digitalización se realiza en las instalaciones de DS-DIGITAL. El archivo es entregado por el cliente sin que DS-DIGITAL tengo acceso a las instalaciones destinadas al archivo físico, por lo que no hay un proceso de organización de documentos en las instalaciones del cliente. El área de ventas también se encarga de la devolución de los archivos físicos al finalizar la digitalización. Al tener estos procedimientos debidamente resueltos, no formarán parte del problema y no serán incluidos en los capítulos posteriores.

Una parte del problema a tomar en cuenta recae en el hecho de que pueden incluirse nuevos tipos de pólizas en cada nueva entrega. Por lo que la solución dada al problema debe considerar cambios en los datos a recolectar, así como los códigos a utilizar para el renombramiento de archivos y la construcción de la estructura determinada para almacenarlos, también el capital humano disponible para realizar dichos cambios. Esto supone el uso de programación modular[18] en los componentes de software que se requieran.

Estado actual del servicio ofrecido por DS-DIGITAL

Actualmente, la aseguradora “A” tiene almacenados en la nube de DS-DIGITAL una gran cantidad de documentos organizados tal cual los ha entregado y está en espera de una solución para las demandas mencionadas anteriormente en este capítulo. Además, el cliente también provee archivos digitales en sus entregas. Es por ello que la solución debe

[18] Programación modular: aquella que usa el concepto de dividir un problema complejo en subproblemas más pequeños, hasta que estos sean fáciles de tratar y resolver por separado.

poder aplicarse a archivos que ya estén en formato digital. Por lo que dar una solución que involucre el proceso de digitalización resulta opcional, aunque deseable.

Sobre la infraestructura de *hardware* de DS-DIGITAL se puede comentar que se encuentra funcionando las 24 horas del día y su servicio de nube se encuentra trabajando en perfectas condiciones. Ese servicio es ofrecido por un tercero: *AWS EC2*[19] - *Google Compute Engine*[20]. Cuenta con suficiente almacenamiento para brindar muchos servicios, empero por razones de confidencialidad no se expresará la capacidad ni especificaciones de dicho servidor en este documento.

Todo el software con el que se trabaja en DS-DIGITAL es de *software* libre, con licencias *GPL*. Únicamente la herramienta para la digitalización es de software privativo: *Capture Perfect 3.0* (Sam Leffler Copyright (c) 1991-1996 *Silicon Graphics*, Inc), por lo que es necesario que las herramientas informáticas desarrolladas para la solución de la problemática planteada estén sujetas a criterios de *software* libre. Esto implica la utilización de lenguajes de programación, librerías y frameworks libres, etc.

[19] *AWS EC2*: servicio ofrecido por *Amazon Inc*. web que provee de seguridad, recursos de cómputo alcanzables y expandibles en la nube. Está diseñado para hacer una web escalable, y fácil para desarrolladores.
[20] *Google Compute Engine*: solución similar a la ofrecida por *AWS EC2* pero plataforma ofrecida por *Google Inc.*

Capítulo III: Sistema inteligente de captura de datos

La solución propuesta para el problema de separación, de reconocimiento óptico de caracteres, de la identificación de las palabras clave, de la creación de la estructura básica y del el renombramiento de pólizas de seguro se denominará con las siglas SICD: Sistema Inteligente de Captura de Datos.

¿Qué será el SICD?

El SICD será una aplicación *web* que va a permitir al cliente resolver la problemática antes planteada:

- Almacenamiento de pólizas: El SICD permitirá subir las pólizas a la plataforma de DS-DIGITAL.
- Separación de las pólizas: Mediante identificadores, el SICD separará en archivos pdf cada una de las pólizas. Los identificadores serán variables.
- *OCR*: El SICD generará archivos de texto con las palabras de cada archivo pdf.
- Identificación de palabras clave: Para cada tipo de archivo se extraerán de una base de datos las claves y las expresiones regulares que permitan identificarlas.
- Estructura básica: Se renombran los archivos conforme al código y se almacenarán en una estructura de directorios. Tanto el código como la estructura de directorios serán generadas a partir de información en la base de datos.

El SICD estará alojado en *GNU/Linux CentOS*[21] *7*, por lo que las herramientas para desarrollar e implantar el SICD deben ser compatibles con los sistemas operativos *Red Hat Enterprise Linux*.

DS-DIGITAL siendo una PyME no se puede permitir adoptar soluciones de *software* costosas, que incrementen el precio de los productos y servicios finales. Tampoco puede

[21] *CentOS:* es una bifurcación a nivel binario de la distribución *Linux Red Hat Enterprise Linux RHEL*, compilado por voluntarios a partir del código fuente publicado por *Red Hat* (*Community Enterprise Operating System*).

hacer uso de *software* de privativo de manera ilegal poniendo en riesgo a la empresa, a sus trabajadores y faltando gravemente a la ética profesional. Es por ello que todas las herramientas de *software* utilizadas deben estar licenciadas con *GPL.*

El SICD debe ser una aplicación modular que permita en el futuro ser adaptada a necesidades de captura de datos mediante *OCR* para otro tipo de archivos físicos y no solamente pólizas de seguro. Sin embargo la capacidad de añadir nuevos módulos a la aplicación se concibe como un desarrollo a futuro, por lo que no se incluirá en este documento.

Soluciones similares en el mercado

A continuación se mencionan algunas herramientas o servicios disponibles en el mercado que enfrentan desafíos similares a los que el SICD se propone resolver.

ABBYY Finereader

Características:

- Edición y comentarios de archivos *PDF.*
- Conversión de archivos *PDF* y otros formatos escaneados.
- Comparación de documentos.
- Conversión automática (*Hot Folder*) hasta 5,000 páginas.

Precios:

- Por las dos primeras características USD$199.00 al mes.
- Por las cuatro características USD$399.00 al mes (Por qué *FineReader*, 2019).

ABBYY es una de las empresas líder en cuanto a tecnologías de la digitalización. Fue pionero en uso de *OCR* en el *software* de escaneo. *ABBYY Finereader* provee una solución factible al problema del SICD, sin embargo **no automatiza el proceso de renombrado ni crea la estructura de directorios.** Para satisfacer las necesidades del

cliente se requeriría trabajo manual de un equipo de capturistas y el costo de producción se incrementaría considerablemente.

SODA PDF Online

Características:

- Convierte archivos en formato de imagen a archivos *PDF* editables.
- Permite convertir archivos *PDF* a otros formatos de oficina como *Word, Excel*, etc.
- Modifica archivos *PDF* con opciones como unir, separar, comprimir, etc.
- Inserta elementos de página.

Precio:

- USD$2.00 por cada página o imagen.
- USD$20.00 por utilizar *OCR* en las conversiones. (*Soda PDF Anywhere*, 2019).

Actualmente existen muchas empresas que abren al público servicios de oficina sencillos como este. Siguiendo el ejemplo de *ABBYY*, ofrecen conversiones y opciones de editado con un costo bajo por volumen. *Soda PDF* es quizá uno de los más conocidos. A pesar de que pueden solucionar algún problema de oficina, las capacidades reales de acción y edición son bastante limitadas y los precios son bastante altos si se trabaja con volúmenes de miles de documentos. Además no resuelve la necesidad de la automatización.

Algunas aplicaciones web similares son:

- ❑ *SmallPDF*
- ❑ *CleverPDF*
- ❑ *Online2PDF*
- ❑ *IcecreamPDFConverter*

Aplicaciones para renombramiento de archivos por lotes

Existen programas que sirven para renombrar archivos en masa. Por ejemplo:

- *Bulk Rename Utility*
- *LupasRename*
- *ReNamer PortableFileRenamer*
- *AdvancedRenamer*

(Moya, 2016)

Todas ellas tienen la característica común de tener versiones gratuitas **que no son libres**. En la mayoría de los casos las aplicaciones vienen con un *adware*[22] que puede comprometer la seguridad del usuario y además infectar al navegador con propaganda no solicitada. En el caso de *FileRenamer* la versión gratuita no contiene *adwares* pero resulta limitada. Además **ninguna de las aplicaciones soporta la automatización de renombramiento con datos** obtenidos mediante *OCR*.

Dokmee Document Manager

Existen empresas que han desarrollado soluciones verdaderamente completas. Una de ellas es *Office Gemini*, ubicada en Houston Texas. Su producto *Dokmee* es un manejador de documentos bastante completo. Lo ofrece en tres presentaciones: escritorio, profesional y para corporativos (*Enterprise*). No tiene versión gratuita aunque permite una prueba de 30 días sin costo. (*Office Gemini, LLC*, 2019).

Dokmee permite buscar palabras, códigos de barras y otros elementos dentro de un *PDF* escaneado. Funciona a través de plantillas que pueden ser personalizadas con una interfaz gráfica en la que mediante una herramienta bastante robusta; permite al

[22] *adware:* programas diseñados para mostrar publicidad en una computadora, redirigir solicitudes de búsqueda a sitios web de publicidad y recopilar datos comerciales sin el consentimiento del usuario.

cliente identificar los elementos en tiempo real, viendo la identificación en una previsualización del documento.

Sin duda una herramienta completísima. Su punto débil es el precio. *Dokmee* cuesta $279.00 dólares mensuales en su versión de escritorio por una sola estación de trabajo y por un tiempo limitado no especificado. Las versiones profesional y corporativa simplemente no muestran un precio al público en su página de internet. Por lo que es de suponer que se requiere pedir una cotización bastante costosa. Para resolver las necesidades que se requieren en el presente proyecto, se requeriría por lo menos de la versión profesional.

Análisis y Diseño

Análisis del problema

Existen 3 problemas que DS-DIGITAL no ha logrado resolver y dar solución a ellos es el objetivo fundamental del SICD:

1. Automatizar la separación de archivos.
2. Automatizar la extracción de palabras clave por *OCR.*
3. Automatizar la creación de una estructura de los archivos renombrados.

Cada uno de estos problemas requiere un análisis particular.

Separación de archivos

En cuanto a la separación de archivos existen tres casos posibles según los documentos que pueden encontrarse alojados en DS-DIGITAL y una excepción a todos. Se iniciará el análisis por dicha excepción.

Se supone que el archivo en *PDF* tiene únicamente imágenes sin texto. Si bien la mayoría de los documentos escaneados suelen ser guardados en un formato que ya tenga incluido texto seleccionable, pueden existir excepciones. Para ello se utilizará la herramienta *OCR* de GNU/Linux *ocrmypdf* con el siguiente comando:

```
ocrmypdf [archivo sin texto] [archivo con texto]
```

Con esto se podrá continuar con los casos generales.

Caso A: El número de hojas de cada documento es fijo, por lo que solo hay que separar por número de hojas.

Caso B: Los documentos ya están separados, por lo que no es necesario volverlos a separar.

Caso C: Los documentos no están separados y tienen un número de hojas variable.

Para atender los tres casos se asignará un código a cada uno que contenga la información clave para resolverlo. Dicho código se guardará en una base de datos y el sistema lo utilizará para resolver el problema de separación en cada caso.

Caso A: Fijo%[número de páginas].

Caso B: Nulo.

Caso C: [Inicio/Fin]]%[clave].

El símbolo % se usa como separador de campos.

[número de páginas] es un número entero que indica cuántas páginas se van a separar.

[clave] es una palabra o expresión regular cuya presencia indique una nueva página.

inicio / fin indican si al encontrar la clave se trata del fin de un documento o del inicio.

Nótese que para conseguir éxito en el caso C el sistema requiere analizar con *OCR* el documento. Para mejorar el desempeño es necesario dejar ese caso como última opción.

Con esta información se puede hacer uso del programa *pdftk* para separar los *PDF* según la relación de las páginas donde termina y donde empieza el documento.

Con el comando :

```
pdftk [nombre del archivo] burst
```

Se separará todo el documento *PDF* en páginas nombradas por su número de página; después con el comando:

```
pdftk [archivo 1] [archivo 2] cat output [archivo unido]
```

Se podrán ir juntando los archivos que correspondan a un solo documento. Con el desarrollo de un *shell script*[23] en *Bash*[24]*,* el cual trabajará con todos los documentos e irá uniendo archivos mediante un arreglo[25] que contenga los números de las páginas de inicio y/o de final de cada póliza. Dicho arreglo puede ser obtenido en el caso A, en que se trata de una progresión uniforme. De tal manera que el *script* puede ser ejecutado desde la plataforma *web*. Para el caso C es necesario un análisis adicional.

Encontrar las páginas donde se encuentra una palabra clave es posible utilizando la herramienta *pdfgrep*, que nos permite realizar búsquedas precisas y eficientes en archivos *PDF*. *Pdfgrep* es una adaptación del poderoso comando *grep* con opciones particularmente útiles si se trabaja con archivos *PDF*.

```
pdfgrep -ni [palabra clave] [archivo]
```

[23] *shell script:* es un programa de computadora diseñado para ser ejecutado por la terminal de comandos de Unix.
[24] *bash:* es un programa informático, cuya función consiste en interpretar órdenes, y un lenguaje de consola (Bourne-again shell).
[25] arreglo*:* estructura de datos que sirve para almacenar grandes secuencias de datos.

Donde:

- La opción -n muestra la página donde la búsqueda tiene éxito.
- La opción -i hace que la búsqueda sea insensible a mayúsculas y minúsculas.
- La palabra clave puede incluir expresiones regulares.
- El archivo puede incluir expresiones regulares para buscar en varios archivos.

Para este caso solo se realizará una búsqueda en un único archivo. Cada número de página en que se encuentre la palabra clave se guardará en el arreglo que se usará para unir las páginas.

Extracción de palabras clave por *OCR*

Para la extracción de palabras clave se tendrá solo un caso general.

Es necesario utilizar de nuevo *pdfgrep* para guardar un arreglo con las palabras clave. Se tendrá que utilizar el comando tantas veces como palabras clave lo necesite.

```
pdfgrep -i [palabra clave] [archivo]
```

En esta ocasión se utilizarán expresiones regulares para abarcar todos los archivos *PDF* pequeños. Los arreglos con las palabras clave nos ayudarán a concluir con el tercer paso.

Creación de la estructura

Finalmente, con los archivos debidamente separados y con las palabras clave reunidas en memoria, se procederá con el renombramiento. La construcción de otro *script* en *Bash* podrá resolver el problema de manera eficaz utilizando comandos esenciales de *GNU/Linux* como *mkdir*, *mv*, etc..

Es importante incluir en el *script* comandos para limpiar los archivos intermedios, de tal forma que no quede rastro en el directorio de trabajo en la tarea que se ha llevado para procesar los documentos.

Diagrama de flujo

A continuación se presenta un diagrama de flujo que resume el procesamiento de los archivos.

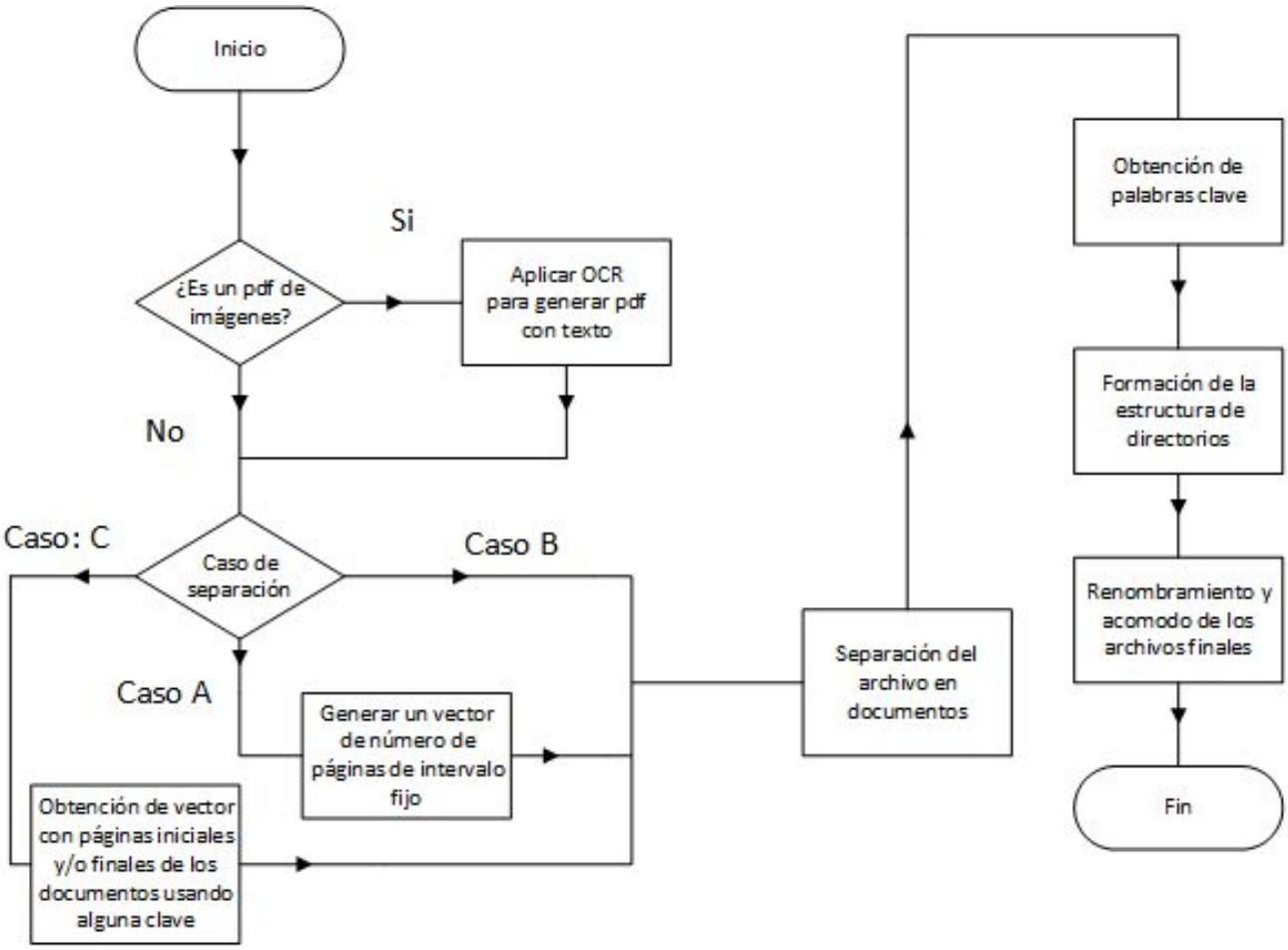

Figura 3.1 Diagrama de flujo del módulo de procesar archivos del SICD.

Tipos de usuarios del SICD

El SICD debe ser utilizable por medio de un portal *web* para sus diferentes usuarios. Debe poder resolver las necesidades requeridas por el cliente, quien sería el usuario final. Sin embargo no es el único usuario, ya que se requiere un usuario administrador y un usuario digitalizador.

Los perfiles de los usuarios son los siguientes:

Usuario final: Cualquier persona cuyo acceso al SICD sea solicitado por el cliente previo acuerdo con el área de ventas de DS-DIGITAL.

Digitalizador: Colaborador de DS-DIGITAL que se encargará del proceso de digitalizar los documentos y procesar el archivo resultante. Puede darse el caso de que solo se encargue de procesar al archivo. Es el usuario que maneja la parte operativa del SICD. Debe ser capaz de modelar expresiones regulares para el reconocimiento de palabras clave, por lo que de preferencia se espera que sea una persona con conocimientos en Informática, Ciencias de la Computación o Ingeniería en Sistemas Computacionales.

Administrador: Persona a cargo del SICD. Se trata de un ejecutivo de DS-DIGITAL. Gestionará los proyectos, los usuarios finales y los digitalizadores.

De momento solo se requerirá de un administrador. En el futuro el SICD podría contar con un usuario llamado “Director de proyecto” que podría ocupar un lugar intermedio entre el administrador y los digitalizadores.

Es importante mencionar que dos de los tres tipos de usuarios laboran en DS-DIGITAL. Esto es debido a que el SICD está pensado para ofrecer un servicio que brinde una solución sencilla al usuario final, pero que debe ser supervisado y mantenido por personal de DS-DIGITAL. La automatización del sistema podría ser un proyecto a futuro si se encuentra un modelo de negocio que así lo requiera.

En la tabla siguiente se describen las capacidades habilitadas en el SICD para cada uno de los usuarios antes listados:

Capacidad	Usuario final (cliente)	Digitalizador	Administrador
Acceso a los archivos a través de la dirección de la nube	Si	No	No
Capacidad para subir archivos	Si	Si	No
Capacidad para generar una estructura de pólizas a partir de un *PDF*	No	Si	No
Capacidad para gestionar tipos de documentos y estructuras	No	Si	No
Capacidad para gestionar usuarios finales	No	No	Si
Capacidad para gestionar digitalizadores	No	No	Si
Tiempo de vida del usuario está definido por :	El contrato con el cliente	El tiempo que dure el proceso de digitalización	El tiempo de vida de la aplicación

Tabla 3.1 Tabla de características de usuarios del SICD.

Ya definidas las capacidades y privilegios que cada usuario tendrá, se continua en establecer las reglas de negocio y los requerimientos funcionales y no funcionales para el desarrollo del SICD.

Reglas de negocio del SICD

ID	Nombre	Descripción
RN01	*Login*	El SICD contará con un *login* obligatorio. El sistema reconocerá el tipo de usuario una vez confirmado el *login*.
RN02	Administración de usuarios	El usuario administrador gestiona[26] a usuarios finales y digitalizadores.
RN03	Administración de proyectos	El usuario administrador gestiona los proyectos y les asocia usuarios, directorios y tipos de documento.
RN03	Subir archivos	El usuario final y el digitalizador pueden subir archivos y crear carpetas.
RN04	Procesar archivos	El usuario digitalizador puede procesar los archivos (separar, generar estructura, capturar palabras clave, etc.).
RN05	Gestión de tipos de documento	El usuario digitalizador gestionará los tipos de documento que se vayan a requerir en cada proyecto.
RN06	Revisión de archivos	El usuario digitalizador revisa los archivos procesados.
RN07	Gestión de archivos	El usuario final podrá subir, bajar, eliminar y consultar archivos en su directorio en la nube.

Tabla 3.2 Tabla de Reglas de negocio del SICD.

[26] Entiéndase gestionar como realizar altas, bajas y cambios.

Los **tipos de documento** son parte importante de la solución al problema. Cada tipo de póliza o documento que se vaya a procesar necesitará de tres elementos:

1. Una lista de palabras clave con expresiones regulares que permitan su extracción.
2. Una manera de identificar unos documentos de otros para la separación.
3. Una forma de estructurar los subdirectorios y los archivos.

Esto es lo que el usuario digitalizador gestionará en la *RN05*. Al referirnos a "tipo de documento" en adelante, se hace referencia a un concepto que encapsula estos tres elementos.

Para que el usuario administrador pueda gestionar de manera adecuada a los usuarios en su ámbito de trabajo se definirá a continuación el concepto de **proyecto** (RN03).

Un proyecto es una solución de un cliente determinado, contiene un directorio de trabajo y al menos un tipo de documento. Así como uno o más usuarios finales.

El proyecto tendrá dos fases: fase de trabajo y fase normal.

Mientras el proyecto esté en *fase de trabajo* tendrá asignado al menos un digitalizador que trabaje en él y sea capaz de procesar archivos. Terminado el trabajo, el proyecto entraría en *fase normal* que consiste únicamente en almacenar los archivos ya procesados.

Cabe mencionar que un cliente puede tener varios proyectos y un proyecto puede encontrarse en fase de trabajo más de una vez.

La solución final al problema planteado se da en la regla de negocio RN03 llamada procesar archivos. Es donde el sistema extrae información, separa el documento, lo renombra con alguna clave y organiza los archivos resultantes con una estructura.

A continuación se listan los requisitos funcionales y no funcionales del sistema.

Requerimientos Funcionales del SICD

Identificación del requerimiento	RF01
Nombre del requerimiento	Autentificación de usuario.
Características	El usuario debe autenticarse con un nombre de usuario y contraseña.
Descripción del requerimiento	El nombre de usuario y la contraseña serán comunicadas al usuario por el administrador.
Usuarios	● Administrador ● Digitalizador ● Usuario Final

Tabla 3.3 Requerimiento funcional RF01.

Identificación del requerimiento	RF02
Nombre del requerimiento	Administración de proyectos.
Características	El administrador gestiona los proyectos (crear, eliminar, editar, etc).
Descripción del requerimiento	Cada proyecto debe tener un identificador y una descripción. Cada proyecto debe tener asociados usuarios finales, digitalizadores, tipos de documentos y un directorio de trabajo.
Usuarios	● Administrador

Tabla 3.4 Requerimiento funcional RF02.

Identificación del requerimiento	RF03
Nombre del requerimiento	Administración de usuarios.
Características	El administrador gestiona los usuarios (crear, eliminar, editar, etc).
Descripción del requerimiento	Cada usuario debe tener un identificador y datos personales, así como definido un tipo (usuario final o digitalizador). Los digitalizadores y los usuarios finales están asociados a uno o más proyectos.
Usuarios	● Administrador

Tabla 3.5 Requerimiento funcional RF03.

Identificación del requerimiento	RF04
Nombre del requerimiento	Gestión de tipos de documento.
Características	El usuario digitalizador gestiona los tipos de archivos (crear, eliminar, editar, etc).
Descripción del requerimiento	Cada tipo de archivo contará con un identificador único por proyecto. Además tendrá un nombre, un código que defina la estructura a utilizar, un código que defina el método de separación de los documentos y una lista de expresiones regulares para la captura de palabras clave. Los tipos de archivos pueden estar asociados a uno o más proyectos.
Usuarios	● Digitalizador

Tabla 3.6 Requerimiento funcional RF04.

Identificación del requerimiento	RF05
Nombre del requerimiento	Subir archivos.
Características	El usuario puede subir archivos y crear directorios.
Descripción del requerimiento	Debe existir una interfaz de gestión de archivos con botones para subir archivo, crear directorios e ir al directorio padre.
Usuarios:	● Usuario Final ● Digitalizador

Tabla 3.7 Requerimiento funcional RF05.

Identificación del requerimiento	RF06
Nombre del requerimiento	Procesar archivo.
Características	El usuario digitalizador puede escoger un archivo para ser procesado. Esto incluye la separación en documentos, la captura automática de palabras clave y el acomodo según la estructura.
Descripción del requerimiento	El usuario digitalizador debe escoger un nombre para el subdirectorio donde se va a generar la estructura. El usuario digitalizador debe escoger el tipo de documento para procesar el archivo elegido.
Usuarios:	● Digitalizador

Tabla 3.8 Requerimiento funcional RF06.

Identificación del requerimiento	RF07
Nombre del requerimiento	Revisar documentos procesados.
Características	El usuario digitalizador podrá revisar los documentos que ya hayan sido procesados con la finalidad de comprobar el desempeño del procesamiento.
Descripción del requerimiento	El usuario digitalizador podrá navegar en una interfaz que permita revisar los archivos procesados. La interfaz debe contar con una vista previa de los archivos.
Usuarios:	● Digitalizador

Tabla 3.9 Requerimiento funcional RF07.

Identificación del requerimiento	RF08
Nombre del requerimiento	Gestión de archivos.
Características	El usuario final puede gestionar los archivos en su directorio de trabajo (descargarlos, visualizarlos, etc.).
Descripción del requerimiento	Requerirá una interfaz para navegar por directorios, así como previsualizar y abrir archivos.
Usuarios:	● Usuario final

Tabla 3.10 Requerimiento funcional RF08.

Requerimientos no funcionales

Identificación del requerimiento	RNF01
Nombre del requerimiento	Gestión de archivos exclusivo del usuario final.
Características	El usuario final es el único que puede bajar y abrir sus documentos.
Descripción del requerimiento	Debe evitarse que el usuario digitalizador pueda abrir o descargar archivos de algún directorio de trabajo.
Usuarios:	● Usuario digitalizador

Tabla 3.11 Requerimiento funcional RNF01.

Identificación del requerimiento	RNF02
Nombre del requerimiento	Asociación *proyecto-digitalizador.*
Características	Los usuarios digitalizadores solo estarán asociados a los proyectos el tiempo estrictamente necesario para cumplir con sus funciones.
Descripción del requerimiento	El administrador es responsable de dar permisos sobre los proyectos a los usuarios digitalizadores solo el tiempo debido.
Usuarios:	● Administrador

Tabla 3.12 Requerimiento no funcional RNF02.

Identificación del requerimiento	RNF03
Nombre del requerimiento	Asociación *proyecto-usuario* final.
Características	Los usuarios finales sólo pueden tener control de los archivos en su directorio de trabajo mientras tengan un contrato vigente con DS-DIGITAL.
Descripción del requerimiento	El administrador es responsable de dar permisos sobre los directorios de trabajo a los usuarios finales solo el tiempo debido.
Usuarios:	● Administrador

Tabla 3.13 Requerimiento no funcional RNF03.

Identificación del requerimiento	RNF04
Nombre del requerimiento	Tipo de documentos.
Características	El usuario final no tendrá acceso a la información que permite procesar los archivos, ni tendrá la capacidad de procesarlos.
Descripción del requerimiento	La información para procesar los archivos quedará almacenada en la nube pero no estará disponible para los usuarios finales. Solo los digitalizadores tendrán acceso y harán uso de dicha información.
Usuarios:	● Administrador

Tabla 3.14 Requerimiento no funcional RNF04.

Casos de uso del SICD

Para evitar una extensión desmedida en el documento sólo se documentan los diagramas de casos de uso general y aquellos casos que requieran un análisis particular debido a su complejidad.

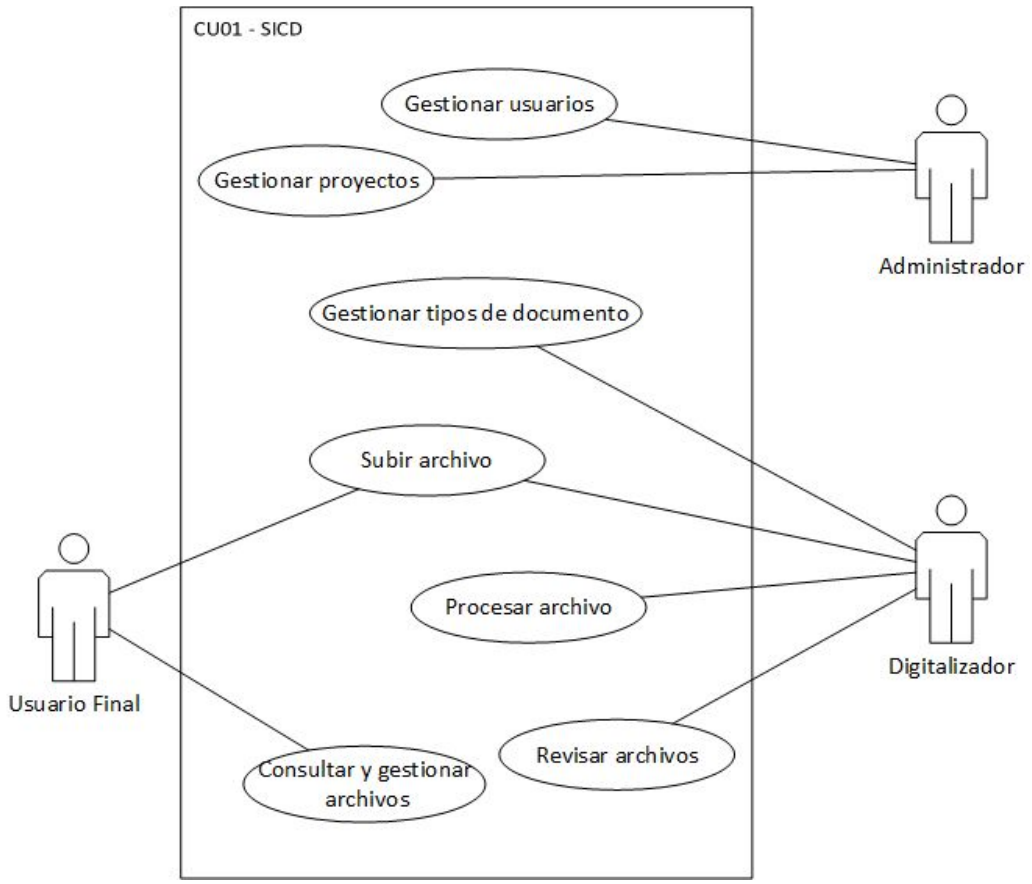

Figura 3.2 Diagrama general de casos de uso del SICD.

Autor:	Cristian Cardoso Arellano.
Nombre del caso de Uso:	Gestionar usuarios.
Propósito:	Proporcionar la capacidad al administrador de gestionar usuarios.
Descripción:	El usuario administrador crea, elimina o edita usuarios.
Precondición:	Para editar o eliminar debe existir al menos un usuario.
Postcondición:	Todos los cambios se verán reflejados en la base de datos del sistema.
Entrada:	1. Selección de usuarios a eliminar o editar (*checkbox*[27]). 2. Datos de los usuarios al crear o editar (nombre, contraseña, etc.).
Referencias:	RF03.
Subcasos de uso:	Crear Usuario, eliminar Usuario, editar Usuario.

Tabla 3.15 CU01-Gestionar usuario.

[27] *Checkbox*: es un elemento de interacción de la interfaz gráfica de usuario (Casilla de verificación).

Autor:	Cristian Cardoso Arellano.
Nombre del caso de Uso:	Gestionar proyectos.
Propósito:	Proporcionar la capacidad al administrador de gestionar proyectos.
Descripción:	El usuario administrador crea, elimina o edita proyectos.
Precondición:	Para editar o eliminar debe existir por lo menos un proyecto.
Postcondición:	Todos los cambios se verán reflejados en la base de datos del sistema.
Entrada:	1. Selección de proyectos a eliminar o editar (*checkbox*). 2. Datos generales del proyecto al crear o editar (nombre, compañía etc.). 3. Lista de usuarios finales y digitalizadores asociados al proyecto. 4. Lista de tipos de datos asociados al proyecto.
Referencias:	RF02.
Subcasos de uso:	Crear proyecto, eliminar proyecto, editar proyecto.

Tabla 3.16 CU02-Gestionar proyectos.

Autor:	Cristian Cardoso Arellano.
Nombre del caso de Uso:	Gestionar tipos de documento.
Propósito:	Proporcionar la capacidad al digitalizador de gestionar tipos de documento.
Descripción:	El usuario digitalizador crea, elimina o edita tipos de documento.
Precondición:	1. Para editar o eliminar debe existir por lo menos un tipo de documento. 2. Debe existir por lo menos un proyecto al que asociar tipos de documento.
Postcondición:	El administrador podrá agregar el tipo de documento creado a los proyectos.
Entrada:	1. Selección de tipos de documento a eliminar o editar (*checkbox*). 2. Datos generales del tipo de documento al crear o editar (nombre, identificador, etc.). 3. Lista de palabras claves. 4. Lista de expresiones regulares asociadas a las palabras clave. 5. Forma de separación de los documentos (caso A, B o C). 6. Si es caso A, número de páginas por documento. 7. Si es caso C palabra clave o expresión regular para reconocer la separación. 8. Si es caso C definir si la palabra clave se va a encontrar en la primera página o la última página.

Referencias:	RF04.
Subcasos de uso:	Crear tipo de documento, eliminar tipo de documento, editar tipo de documento.

Tabla 3.17 CU03-Gestionar tipos de documento.

Autor:	Cristian Cardoso Arellano.
Nombre del caso de Uso:	Subir archivo.
Propósito:	Proporcionar la capacidad al usuario final y el digitalizador de subir archivos.
Descripción:	El usuario final o el digitalizador suben archivos *PDF*.
Precondición:	El usuario final y el digitalizador deben estar asociados a un proyecto.
Postcondición:	El archivo será accesible desde el directorio de trabajo del proyecto.
Entrada:	1. Selección de un archivo.
Referencias:	RF05.
Subcasos de uso:	No aplica.

Tabla 3.18 CU04-Subir archivo.

Autor:	Cristian Cardoso Arellano
Nombre del caso de Uso:	Procesar archivo.
Propósito:	Proporcionar la capacidad al digitalizador de procesar los archivos.
Descripción:	El digitalizador procesa el archivo, genera la estructura, extrae palabras clave, etc..
Precondición:	Debe existir un archivo *PDF* y el usuario digitalizador asociado debe estar asociado al proyecto.
Postcondición:	La estructura de los archivos será visible para los digitalizadores y los usuarios finales en el directorio de trabajo del proyecto en cuestión.
Entrada:	1. Tipo de documento a utilizar (*combo box*[28]). 2. Nombre del directorio raíz de la estructura.
Referencias:	RF06.
Subcasos de uso:	No aplica.

Tabla 3.19 CU05-Procesar archivo.

[28] *Combo box*: es un elemento de interacción de la interfaz gráfica de usuario el cual lista varias opciones para su selección (Cuadro combinado).

Autor:	Cristian Cardoso Arellano
Nombre del caso de Uso:	Revisar archivos.
Propósito:	Proporcionar la capacidad al digitalizador de navegar dentro del directorio de trabajo de un proyecto y previsualizar archivos (sin opción a descargar).
Descripción:	El digitalizador puede revisar los archivos procesados.
Precondición:	Deben existir archivos en el directorio de trabajo del proyecto al que está asociado el digitalizador.
Postcondición:	No aplica.
Entrada:	No aplica.
Referencias:	RF08, RFN01.
Subcasos de uso:	No aplica.

Tabla 3.20 CU06- Revisar archivos.

Autor:	Cristian Cardoso Arellano.
Nombre del caso de Uso:	Consultar y gestionar archivos.
Propósito:	Proporcionar la capacidad al digitalizador de navegar dentro del directorio de trabajo de un proyecto y previsualizar, eliminar, y descargar archivos, además de cambiar el nombre.
Descripción:	El usuario final podrá gestionar los archivos del directorio de trabajo de un proyecto.
Precondición:	No aplica.
Postcondición:	No aplica.
Entrada:	No aplica.
Referencias:	RF08.
Subcasos de uso:	No aplica.

Tabla 3.21 CU07- Consultar y Gestionar archivos.

Los casos de uso particulares del SICD son los siguientes:

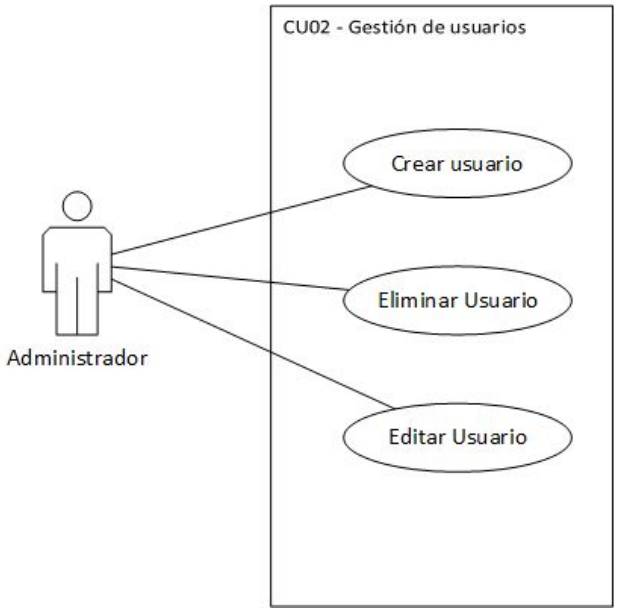

Figura 3.3 Diagrama de casos de uso "CU-02 Gestionar usuarios".

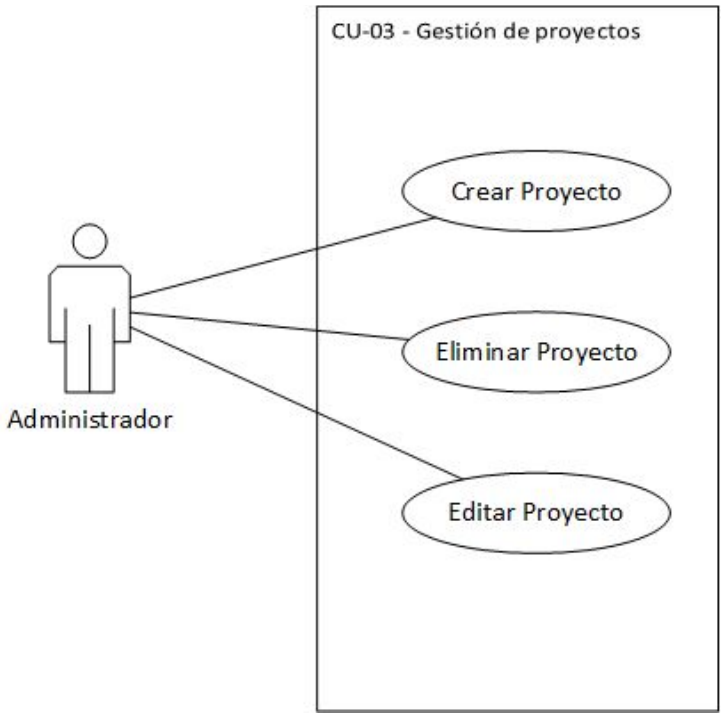

Figura 3.4 Diagrama de casos de uso "CU-03 Gestionar proyectos".

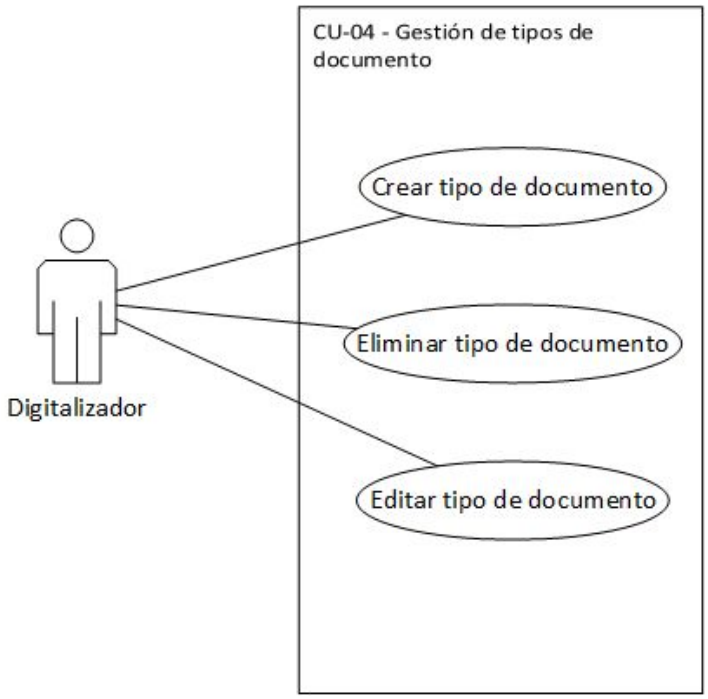

Figura 3.5 Diagrama de casos de uso "CU-04 Gestión de tipos de documento".

En los tres casos se requiere de un formulario para editar y crear, además de una ventana de navegación. Se requiere tener una lista con todos los usuarios, proyectos y tipos de archivo para seleccionar aquellos que se quieran editar o eliminar.

Capítulo IV: Resultados esperados de la implementación del SICD en DS-DIGITAL

Un capturista tarda en promedio 30 segundos en capturar los datos de una o dos póliza en renombrar el archivo de acuerdo a la información capturada. Si un archivo grande consta de más de mil pólizas, puede ser el trabajo de tres días o más. Suma un promedio de 24 horas (tres días laborables). Con el SICD esta operación es casi instantánea, hasta 4,000 pólizas renombradas en 5 minutos.

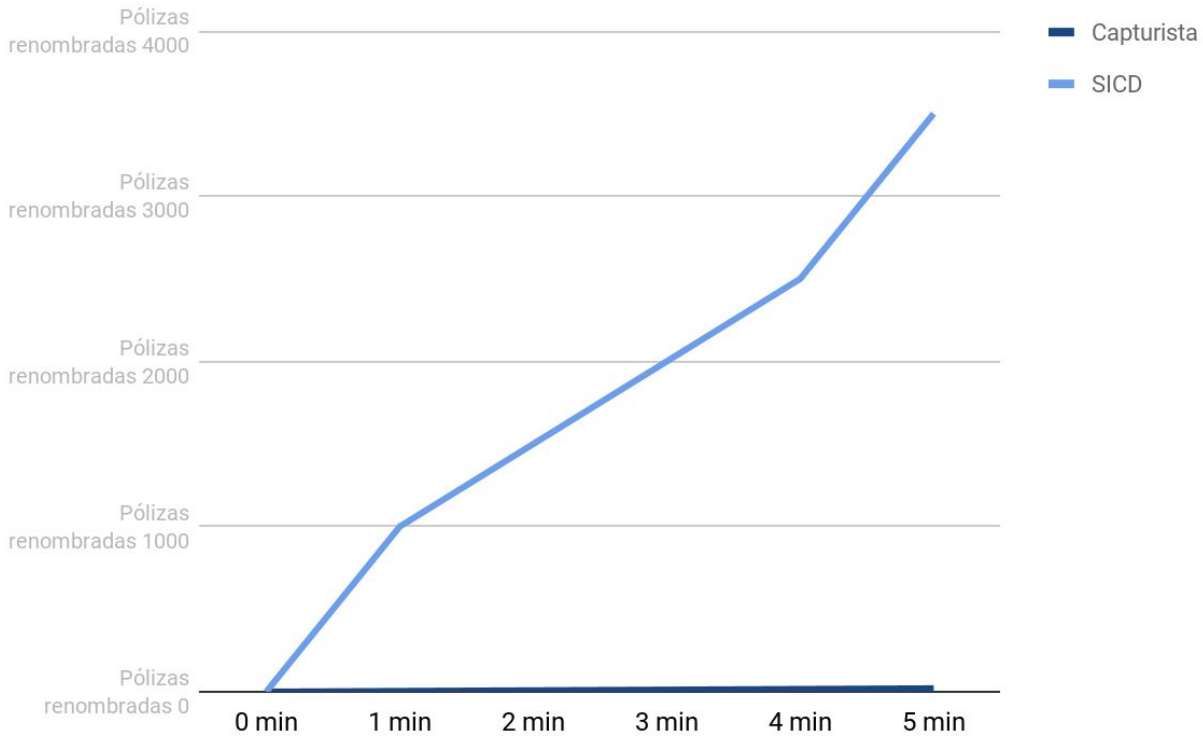

Figura 4.1 Gráfica de tiempo. Capturista vs SICD.

Debido a que el proceso de implementación de expresiones regulares esto podría requerir varias pruebas antes de dar un resultado satisfactorio, se estima que una plantilla sería procesado en lo máximo cinco intentos. Cada intento debe llevar entre 5 y 20 segundos en cuanto a la captura de los datos y el procesamiento de todo el documento,

por lo que en promedio se espera que un archivo con más de 4,000 pólizas esté procesado y listo en 5 minutos o menos, reduciendo de 24 horas (3,600 minutos) a únicamente 5 minutos. Esto supone una reducción del 90% del tiempo empleado en el subproceso de captura de datos y renombramiento de archivos.

Este incremento en la eficiencia se verá reflejado en dos de los servicios ofrecidos por DS-DIGITAL: la organización y el renombramiento de archivos digitales y la digitalización de archivo. Ello implica una reducción en el costo de captura, y que en el costo total del proyecto podría ascender hasta en un 60%.

Conclusiones

Factibilidad del sistema

El diseño del SICD demuestra que el proyecto es factible y perfectamente realizable dadas las características de infraestructura con las que cuenta DS-DIGITAL. Lo mismo se puede decir en cuanto al modelo de negocio seguido por dicha empresa.

El concepto de la captura de datos puede ser en el futuro un referente de DS-DIGITAL que le permita abordar nuevos mercados y expandir el abanico de servicios que ofrece.

Utilidad del software libre

Es de resaltar la gran aplicabilidad del *software* libre en el ámbito empresarial. Muchas veces las empresas optan por soluciones que ya están implementadas y que tienen como características ser de uso limitado, ser costosas y estar diseñadas con propósitos generales. Sin embargo el diseño del SICD ha demostrado que con esfuerzo y conocimientos de sistemas operativos *GNU/Linux*, se pueden utilizar las herramientas de *software* libre para dar soluciones a problemas específicos.

La aplicabilidad del *software* libre en el ámbito empresarial es uno de los valores de DS-DIGITAL y es lo que podría marcar un diferenciador a futuro para destacar en el mercado de las soluciones digitales.

El SICD podría ser la prueba del uso del *software* libre para generar productos de calidad que desarrollen soluciones a bajo costo.

El poder de las expresiones regulares

Las expresiones regulares son una herramienta básica en la filología y son aplicables para el resto de las disciplinas y ciencias del ámbito humano. El poder que tienen para el reconocimiento de patrones es innegable, en especial en el rubro de cadenas de texto.

En el desarrollo del reconocimiento de patrones y el denominado *Machine Learning*[29] se encuentran algoritmos que destacan por su gran capacidad para dicha tarea. Empero también suelen ser algoritmos complejos que requieren un elevado costo de recursos computacionales.

Por su parte las expresiones regulares tienen la capacidad de reconocer casi cualquier código de caracteres con palabras o claves yuxtapuestas. Es por ello que con conocimientos de codificación los resultados de las expresiones regulares brindan soluciones de poco costo computacional a los problemas comunes relacionadas el manejo de archivos digitales.

Los programas de línea de comandos

Dado que no muchas personas en el ámbito laboral manejan la línea de comandos de *GNU/Linux y UNIX* muchas de las herramientas que se encuentran disponibles en este formato pasan desapercibidas por los profesionales de TI en empresas de México.

Es bien sabido que todos los lenguajes de programación de alto nivel utilizados para hacer el *Back-End*[30] tienen capacidades que permiten la comunicación con los sistemas operativos y ejecutar instrucciones en la línea de comandos. De esta forma concebimos la creación personalizada de herramientas prácticas.

Finalmente el diseño del SICD es una muestra de procesos complejos con la interacción del sistema operativo *GNU/Linux*, pues ofrece una gran adaptabilidad, escalabilidad y multiprocesos que se pueden conseguir a futuro con cualquier centro de datos que cuente con una infraestructura adecuada, o bien el cómputo en la nube como lo ofrece *AWS EC2 y Google Compute Engine*.

[29] *Machine Learning :* es una aplicación de inteligencia artificial (AI) que proporciona a los sistemas la capacidad de aprender y mejorar automáticamente a partir de la experiencia sin ser programado explícitamente.
[30] *Back-End:* es el código que se ejecuta en el servidor, que recibe solicitudes de clientes y contiene la lógica para enviar los datos apropiados al cliente (Sistemas posteriores).

Anexos

Anexo A

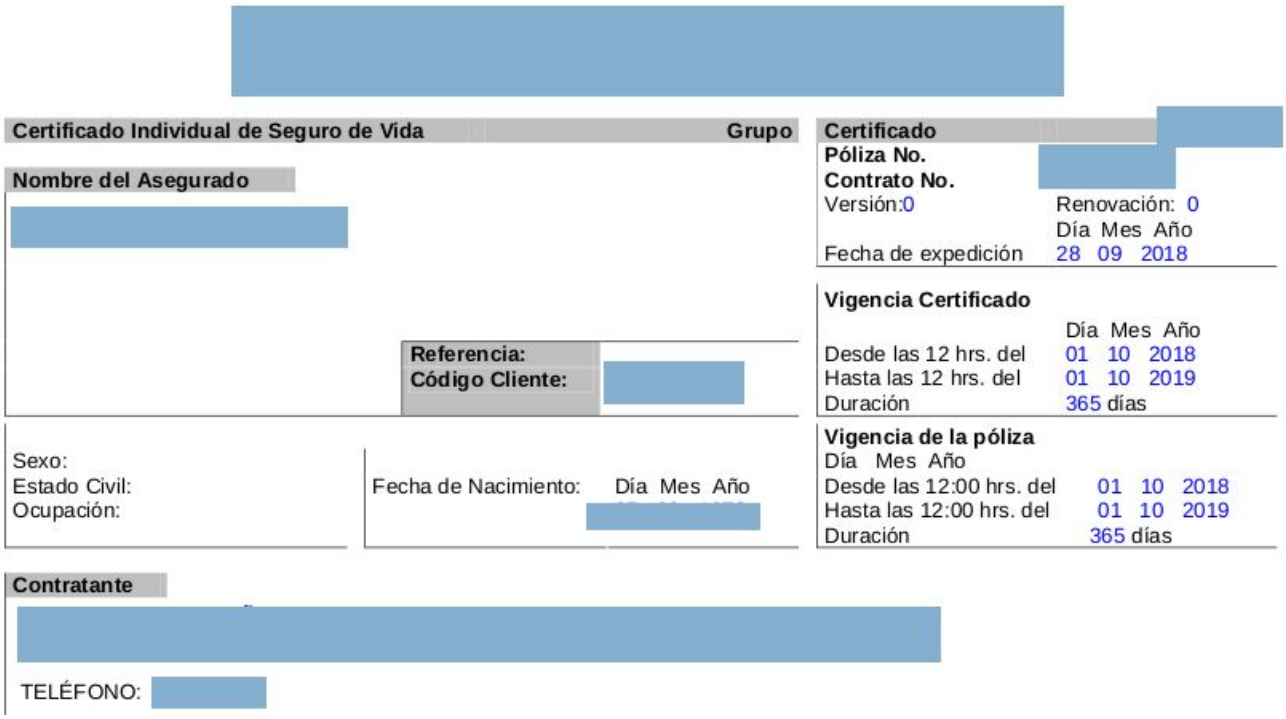

Figura 5.1 Ejemplo de póliza de seguro de vida.

Referencias

Archivo general de la nación. (2015). Recomendaciones para proyectos de digitalización de documentos. Recuperado de https://www.gob.mx/cms/uploads/attachment/file/146401/Recomendaciones_para_proyectos_de_digitalizacion_de_documentos.pdf

Bastin, R., Hurtaud, S., & Senequier, L. (2014, octubre 2). Digitisation of documents and legal archiving. Recuperado de https://www2.deloitte.com/content/dam/Deloitte/lu/Documents/technology/lu_digitisation-documents-legal-archiving_02102014.pdf

Digitalización de documentos: ¿en qué consiste el proceso? (s. f.). Recuperado 3 de enero de 2018, de https://www.ticportal.es/temas/sistema-gestion-documental/digitalizacion-de-documentos

DS-DIGITAL. (2018). DS-DIGITAL: Digitalización, Resguardo y Almacenamiento. Recuperado el 28 de diciembre de 2018, de https://www.facebook.com/dsdigitalmx/

EKCIT. (2018). Digitalización de documentos: ¿en qué consiste el proceso? Recuperado 3 de enero de 2018, de https://www.ticportal.es/temas/sistema-gestion-documental/digitalizacion-de-documentos

Facultad de Contaduría y Administración De la UAEM. (2015, septiembre). Diseño de Procedimientos Precisos de Entrada de Datos. Recuperado de http://ri.uaemex.mx/oca/view/20.500.11799/34538/1/secme-18609.pdf

Grupo Telecon TBS. (2015). Qué es la indexación de los metadatos en la gestión documental? | TBS-Telecon. Recuperado el 27 de diciembre de 2018, de http://www.tbs-telecon.es/blog/que-es-indexacion-metadatos-gestion-documental

ITZELI. (2015, septiembre). Diseño de Procedimientos Precisos de Entrada de Datos. Facultad de Contaduría y Administración de la UAEM. Recuperado de http://ri.uaemex.mx/oca/view/20.500.11799/34538/1/secme-18609.pdf

Moya, P. (2016, mayo 20). Utilidades para renombrar archivos en masa. Recuperado 20 de enero de 2018, de https://omicrono.elespanol.com/2016/05/renombrar-archivos/

Muñoz, M. T. B. (2006, marzo 23). GUIA PARA DIGITALIZAR DOCUMENTOS. Recuperado de http://www.informacionpublicapgr.gob.sv/descargables/sia/academia-de-archivo/guia_digitalizar_documentos.pdf

ONCE. (2016, enero 4). Pautas básicas (en el) uso de OCR. Recuperado de ftp://ftp.once.es/pub/utt/bibliotecnia/Accesibilidad/OCR/Pautas_uso_OCR.pdf

Por qué FineReader. (2019). Recuperado 20 de enero de 2018, de https://www.abbyy.com/es-es/finereader/why-finereader/

Rivas, M. R. (2013, octubre 7). Los eficaces son los nuevos sabios • Forbes México. Recuperado el 28 de diciembre de 2018, de https://www.forbes.com.mx/los-eficaces-son-los-nuevos-sabios/

Soda PDF Anywhere. (2019). Recuperado 20 de enero de 2018, de https://www.sodapdf.com/es/buy/freeonlinetools/dw-success/

Utilidades para renombrar archivos en masa. (2016, mayo 20). Recuperado 20 de enero de 2018, de https://omicrono.elespanol.com/2016/05/renombrar-archivos/

Wisconsin Historical Society. (2018, agosto 8). Digitization Project Guidance For State Agencies. Recuperado de https://www.wisconsinhistory.org/pdfs/la/Digitization-State/State-Digitization-Guidance-Complete.pdf