UNIVERSIDAD DE LAS CIENCIAS INFORMÁTICAS

FACULTAD 3

Grupo de Investigación de Web Semántica

Componente para la extracción automática de metadatos bibliográficos a partir de corpus textuales en formato PDF

Trabajo de Diploma para optar por el título de

Ingeniero en Ciencias Informáticas

Autor:

Leduan Flores Riera

Tutores:

MSc. Yusniel Hidalgo Delgado

Ing. Ernesto Ortiz Muñoz

La Habana, junio de 2016

"Año 58 de la Revolución"

DECLARACIÓN DE AUTORÍA

Declaro ser el autor de la presente tesis y reconozco a la Universidad de las Ciencias Informáticas los derechos patrimoniales de la misma, con carácter exclusivo.

Para que así conste firmo la presente a los ____ días del mes de _____ del año _____.

_____

Leduan Flores Riera

Autor

_____ _____

MSc. Yusniel Hidalgo Delgado **Ing. Ernesto Ortiz Muñoz**

Tutor Tutor

DATOS DE CONTACTO

Síntesis del tutor

El Máster en Ciencias Yusniel Hidalgo Delgado se graduó con Título de Oro en la Universidad de Ciencias Informáticas en el año 2010. En su primer año de adiestramiento desempeñó diversos roles dentro del proyecto de desarrollo del ERP cubano. Actualmente se desempeña como profesor asistente del departamento docente de técnicas de programación de la Facultad 3. Es coordinador del grupo de investigación de Web Semántica de la UCI. Es miembro de la Asociación Cubana de Reconocimiento de Patrones, de la Sociedad Cubana de Matemática y Computación y de la *International Association for Pattern Recognition*.

A mi madre que por su apoyo incondicional y sacrificio he logrado llegar hasta aquí.

A mi padre.

A mi tía y hermanos.

Gracias mamá, por tu cariño y amor, por tu apoyo y consejos, sin ellos no lo hubiera logrado.

Gracias a mi padre por la ayuda y apoyo desde la distancia.

Gracias a mi tía, mi segunda mamá.

Gracias a mis hermanos.

Gracias a mis hermanos de la UCI, Mojena y Pavel.

Al tutor, Yusniel, gracias por los consejos y enseñarme que siempre hay que sacrificarse un poquito más para lograr que el trabajo tenga éxito.

A los compañeros de la brigada, a la gente del grupo de Web Semántica.

RESUMEN

El avance tecnológico y la rápida creación de documentos digitales han permitido el desarrollo de las bibliotecas digitales. Estas se encargan de la gestión documental de los recursos digitales que almacenan, realizando tres procesos fundamentales: la selección, tratamiento y explotación de los recursos. Una de las tareas del tratamiento es la extracción de los metadatos, con el fin de facilitar su explotación, o sea, permitir la búsqueda, acceso y recuperación de la información. La extracción de metadatos es un proceso complejo y costoso, que requiere mucho tiempo y personal altamente calificado para su ejecución, por lo que es necesario contar con herramientas automatizadas que apoyen esta actividad. En el presente trabajo se hace un análisis de tres herramientas implementadas para efectuar la extracción de metadatos automáticamente, además de los métodos que utilizan para la extracción y análisis de la estructura del documento y se lleva a cabo un estudio sobre los lenguajes y estándares que emplean para representar los metadatos. En la investigación se propone un componente web para la extracción automática de metadatos bibliográficos. El componente está basado en tres procesos fundamentales que siguen un flujo de datos representando tuberías y filtros, donde la salida de un proceso constituye la entrada al próximo. Este componente será integrado al proyecto de investigación *"Extracción, publicación y consumo de metadatos bibliográficos como datos enlazados en la web"*, como parte de la fase de *Extracción de Metadatos Bibliográficos,* perteneciente al grupo de investigación de Web Semántica.

Palabras claves: Aprendizaje Automático, Artículos científicos, Bibliotecas digitales, Extracción de metadatos, Documentos PDF, Metadatos, Web Semántica.

ABSTRACT

Technological progress and the fast creation of digital documents has enabled the development of digital libraries. Digital libraries are responsible for management of stored digitals resources and perform three fundamental processes selection, treatment and exploitation of resources. One of the functions of treatment is the extraction of the metadata, in order to facilitate its use, that is, allow the search, access and retrieval of information. Metadata extraction is a complex and expensive process, requiring long and highly qualified staff to run, so it is necessary to have automated tools to support this activity. In this paper is made an analysis of three tools implemented to perform extraction automatically, in addition to the methods used for the extraction and analysis of the document structure and takes place a study of the languages and standards that applications use to display the metadata. In this research is proposed a web component for the automatic extraction of bibliographic metadata. This component is based on three fundamental processes that follow a data stream representing pipes and filters where the output of one process is the input to the next. This component will be integrated into the research project "Extraction, publication and consumption of bibliographic metadata as linked data on the Web", as part of the phase Extraction bibliographic metadata, belonging to the Semantic Web research group.

Keywords: *Machine Learning, Scientific articles, Digital Libraries, Metadata extraction, PDF Documents, Metadata, Semantic Web.*

ÍNDICE

ÍNDICE

ÍNDICE DE TABLAS

ÍNDICE DE FIGURAS

INTRODUCCIÓN

Los avances en el diseño y fabricación de computadoras trajeron consigo el desarrollo de programas de computación más complejos, aparejado a una mayor capacidad de almacenamiento y procesamiento de los datos. Con las mejoras en el hardware y el software se hace necesario compartir información y establecer comunicaciones entre los usuarios. En este contexto, surge Internet como un conjunto de estándares y protocolos. La Internet provee la plataforma y las tecnologías para la creación de la *World Wide Web* (WWW), o Web, por Timothy (Tim) Berners-Lee, para la publicación de contenidos en el formato *Hypertext Markup Language* (HTML).

La Web, desde su creación hasta la actualidad, ha tenido un proceso de evolución y desarrollo con el objetivo de aumentar la calidad con la que se publican los contenidos y ampliar sus funcionalidades y aplicaciones. La Web 2.0 (Web actual) está relacionada con la creación de contenidos dinámicos donde los usuarios crean y comparten la información. Esta tiene las limitaciones del formato, la integración y la recuperación de la información (Hidalgo Delgado y Rodríguez Puente 2013).

Con el objetivo de resolver las limitaciones de la web actual surge la Web Semántica como una extensión de la Web actual. Tim Berners-Lee, promotor del concepto de Web Semántica propone: *"La Web Semántica no pretende sustituir la Web actual, sino que es una extensión de la misma en la que la información tiene un significado bien definido, posibilitando a los humanos y las computadoras trabajar en cooperación"* (Berners-Lee, Hendler y Lassila 2001). A pesar de no estar generalizada debido en gran parte al poco desarrollo de las tecnologías existentes, tiene varias aplicaciones entre las que se encuentran la gestión de documentos digitales y la gestión de referencias bibliográficas (Hidalgo Delgado y Rodríguez Puente 2013).

Las Bibliotecas Digitales son sistemas para la gestión de documentos digitales, donde los usuarios pueden tener acceso a la información desde su ordenador. Las bibliotecas digitales son el resultado de los avances en las tecnologías, de la proliferación de diferentes fuentes de acceso y formatos en los que puede estar la información. Tienen como objetivo lograr el procesamiento, la distribución y explotación de los recursos que almacenan. La biblioteca digital proporciona ventajas para la educación y la investigación debido a la constante actualización de su contenido y a la rapidez con que se puede acceder a ellas.

La biblioteca digital es una evolución de las bibliotecas clásicas donde, en esta última, todo el contenido se encuentra en soporte físico, logrando acceder a él mediante referencias bibliográficas almacenadas en catálogos. Las bibliotecas electrónicas o pequeños repositorios, solo almacenan texto y en algunos casos digitalizan los catálogos de las bibliotecas. La diferencia entre la biblioteca digital y la biblioteca

electrónica está en que las primeras utilizan las telecomunicaciones, es decir, el acceso a la información puede ser realizado de forma remota e independientemente del lugar y el número de conexiones (Hípola, Vargas-Quesada y A. Senso 2000).

Las bibliotecas digitales constituyen un paso de avance en el acceso a la información y al conocimiento, no solo pueden almacenar documentos de texto en varios formatos sino también imágenes, videos, datos, gráficos, audio, figuras 3D y otros. El acceso a la información que almacenan está siempre disponible al estar, las bibliotecas digitales, publicadas en la Web. Al estar disponibles en línea se puede obtener el conocimiento o la información que se busca con rapidez desde el lugar de trabajo o el hogar.

Un ejemplo de documento de texto en una biblioteca digital son los de carácter científico y pueden estar en formatos digitales de texto como PDF, TXT o Word. Los documentos científicos pueden ser: los artículos científicos, las tesis de pregrado, maestrías y doctorados. Cada uno de ellos tiene una estructura definida que se debe tener en cuenta a la hora de su elaboración.

Los artículos científicos deben contener, entre otros elementos: título, autores, resumen, referencias y citas bibliográficas. Estos elementos son atributos que identificarán a cada artículo dentro de la biblioteca digital. Los atributos que identifican a un artículo científico son conocidos como metadatos, los cuales se pueden definir como toda información que describe a un dato o un recurso, dígase artículos científicos, para facilitar su recuperación, autentificación, evaluación, preservación o interoperatividad (Senso y de la Rosa Piñero 2003).

Los metadatos se obtienen a partir de la extracción automática de los mismos, que se define como la extracción de las etiquetas o atributos desde documentos legibles, escritos generalmente en lenguaje natural, por una computadora (Pinilla, Gutiérrez y Ballejos 2014). Para obtener los metadatos desde los artículos científicos se aplican métodos de Aprendizaje Automático (del inglés, Machine Learning), definido como los métodos que se utilizan para representar el proceso de aprendizaje humano en un lenguaje entendible por las computadoras (Mitchell 1997).

Las herramientas para la extracción automática de metadatos bibliográficos pueden ser utilizadas a través de la Web o instalando una aplicación de escritorio. Estas no solo se dedican a la extracción de metadatos, sino que también pueden funcionar como un repositorio personal de artículos científicos o para manejar referencias o citas bibliográficas. Tienen como principal ventaja la rapidez con que son extraídos los metadatos.

En el grupo de investigación de Web Semántica de la Universidad de las Ciencias Informáticas (UCI) se trabaja en el desarrollo de tecnologías basadas en la web semántica para su aplicación en entornos y problemas reales. El grupo de investigación está desarrollando el proyecto de investigación "Extracción, publicación y consumo de metadatos bibliográficos como datos enlazados" y tiene como objetivos extraer, publicar y consumir metadatos bibliográficos siguiendo los principios de los datos enlazados. Una etapa importante en el proyecto es la extracción de metadatos bibliográficos a partir de artículos científicos en formato PDF.

La etapa de extracción de metadatos tiene el objetivo de procesar cada uno de los documentos científicos para obtener sus metadatos bibliográficos. Los metadatos obtenidos en este proceso son el título, los autores, las afiliaciones de cada autor, el resumen y las palabras claves, pertenecientes a la portada de los documentos científicos. Una vez almacenados los metadatos, estos son utilizados en fases posteriores del proyecto de investigación.

La extracción de metadatos requiere de personal altamente calificado para identificar y extraer los metadatos como son el título y los autores, para luego ser guardados en una base de datos en línea (Flynn 2014). La diferencia existente entre la cantidad de documentos digitales a ser procesados manualmente y el número de especialistas que pudieran extraer los metadatos, genera un "cuello de botella" en la realización de este proceso (Khoo, Park y Xia 2009; Liddy et al. 2001).

Como ya se planteó anteriormente, realizar la extracción de metadatos manualmente puede ser muy costoso en cuanto al tiempo. El tiempo real que demora este proceso varía según el dominio que tenga un especialista en realizar el proceso y el propósito por el cual son extraídos los metadatos, (Sicilia 2014). Por ejemplo, el tiempo de archivado de los metadatos de un artículo en un repositorio institucional se ha estimado que demora 5 minutos y 37 segundos como promedio por cada uno de los documentos (Carr y Harnad 2005).

De la problemática descrita anteriormente se deriva el siguiente **problema a resolver**:

¿Cómo *reducir el tiempo empleado por los especialistas en bibliotecología* para *la extracción de metadatos bibliográficos a partir de documentos en formato PDF*?

Objeto de estudio: Extracción de metadatos

Campo de acción: Extracción automática de metadatos de documentos en formato PDF utilizando técnicas de aprendizaje automático.

Para resolver el problema se identifica el siguiente **objetivo general**:

Desarrollar un componente para la extracción de metadatos bibliográficos a partir de documentos en formato PDF utilizando técnicas de aprendizaje automático.

A partir de lo planteado anteriormente se desglosan los siguientes **objetivos específicos**:

1. Elaborar el marco teórico y el estado del arte del objeto de estudio de la investigación mediante el análisis bibliográfico documental para identificar tendencias y adoptar posiciones al respecto.

2. Diseñar una aplicación informática para la extracción de metadatos bibliográficos a partir de documentos en formato PDF utilizando un enfoque de aprendizaje automático.

3. Implementar una aplicación informática para la extracción de metadatos bibliográficos a partir de documentos en formato PDF utilizando un enfoque de aprendizaje automático.

4. Validar los resultados obtenidos con la utilización de la aplicación informática desarrollada mediante la realización de un diseño experimental.

Se tiene como **idea a defender**:

Si se desarrolla un componente para la extracción de metadatos bibliográficos a partir de documentos en formato PDF utilizando técnicas de aprendizaje automático, entonces se reducirá el tiempo empleado para el proceso de extracción por los especialistas en bibliotecología.

Durante la investigación se han empleado un conjunto de **métodos científicos** para el análisis y obtención de la información relacionada con el objeto de estudio y campo de acción de la investigación, los cuales se explican a continuación.

Métodos Teóricos

Con la aplicación del método **Analítico-Sintético**, se analizaron cada uno de los conceptos, que de una forma u otra guardan relación con el objeto de estudio de la investigación y las relaciones existentes entre ellos. Con los resultados de este análisis se realiza una síntesis para comprender las características del objeto de estudio y sus relaciones. Este método permitió definir el marco teórico de la investigación propuesto en el Capítulo 1.

Se utilizó el método **Inductivo-Deductivo** en la confección del estado del arte de la investigación. Este método se basa en los procedimientos inducción y deducción, donde el primero permite arribar a conclusiones generales a partir del estudio de las aproximaciones existentes sobre el objeto de la

investigación y el segundo posibilita a partir de un razonamiento lógico sobre lo anterior inferir nuevos conocimientos que lleven al planteamiento de una posible solución para la investigación.

Métodos Empíricos

Entre los métodos empíricos existentes se seleccionaron para su utilización el de **Medición** y el **Experimental**, el primero para obtener información numérica acerca del tiempo que se demora un especialista en bibliotecología y el componente desarrollado en el proceso de extracción de metadatos. Para la validación de la solución propuesta se hace uso del método **Experimental**, tomando como muestra los datos obtenidos en la medición.

El presente trabajo de diploma consta de una introducción y de tres capítulos, a continuación, un resumen de cada uno de ellos:

En el ***Capítulo 1. FUNDAMENTACIÓN TEÓRICA***: se realiza un análisis de la literatura consultada, se definen los principales conceptos asociados al área del conocimiento en cuestión, que sirvieron de apoyo a la investigación y específicamente los conceptos sobre la extracción automática de metadatos. Como elemento medular del capítulo se hace una exposición y comparación de las herramientas existentes para la extracción automática de metadatos bibliográficos. Además, se explican algunos de los métodos utilizados para la extracción de los metadatos, los estándares y lenguajes para su representación.

El ***Capítulo 2. DESCRIPCIÓN DE LA PROPUESTA****:* está compuesto por la presentación de la propuesta de solución y de una descripción de la misma, a través del análisis del flujo de datos y la caracterización de los procesos fundamentales que la componen. Además, se describe la arquitectura y sus componentes, incluyendo el conjunto de productos de trabajos generados durante la etapa de diseño e implementación de la solución.

El ***Capítulo 3. VALIDACIÓN DE LA PROPUESTA****:* tiene como elemento central la validación de la solución propuesta a partir de las pruebas de software que propone la metodología de desarrollo AUP-UCI. Se explica el diseño experimental utilizado para demostrar si se resuelve o no el problema descrito a partir de un conjunto de datos. La propuesta de solución es validada a partir del diseño experimental, el cual determina si se reduce el tiempo de extracción de los metadatos bibliográficos que emplean los especialistas en bibliotecología.

CAPÍTULO 1. FUNDAMENTACIÓN TEÓRICA

1.1. Introducción

En el presente capítulo se estudian las bibliotecas digitales, el concepto de metadatos, y se define el aprendizaje automático. Se identifican y caracterizan los métodos de aprendizaje automático y las herramientas existentes para la extracción automática de metadatos desde corpus textuales en formato PDF. Se establece una comparación entre las herramientas seleccionadas para el análisis del estado del arte de la investigación. Se describen los lenguajes y estándares para la representación de metadatos.

1.2. Análisis bibliométrico y documental

En la presente investigación el análisis bibliométrico documental se realiza con el objetivo de mostrar la novedad de la revisión bibliográfica realizada, a partir del análisis de las fechas de las publicaciones consultadas. La base de datos de consulta utilizada para la búsqueda de publicaciones fue *Google Scholar*. Las fuentes bibliográficas utilizadas en la investigación son: artículos de revistas, libros, tesis, artículos en conferencias. El análisis realizado se muestra en la siguiente tabla.

Tabla 1: Análisis bibliométrico y documental

Tipo de fuente bibliográfica	Cantidad consultada	Cantidad publicada en los últimos cinco años (2011-2016)
Libros	4	2
Artículos de revistas	30	16
Tesis	1	1
Artículos en conferencia	7	3
Otros documentos	2	1
Total	44	23

Los resultados obtenidos en la **Tabla 1** son representados en el gráfico siguiente, donde se muestra que para un 52% de la bibliografía consultada es del período entre los años 2011 y 2016.

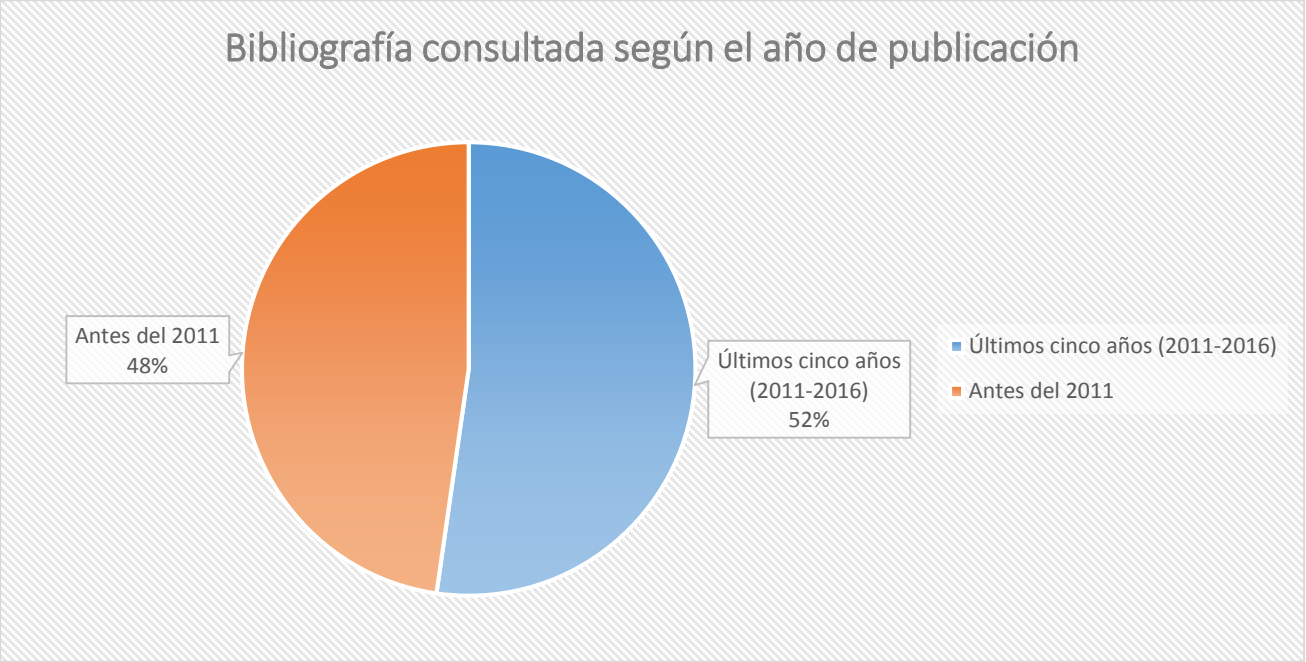


Figura 1: Actualidad de la bibliografía consultada

1.3. Marco teórico

En la década de los 90s creció el interés académico y profesional en el uso de las ***bibliotecas digitales***. En este período se almacenan nuevas revistas en línea e impresas en bibliotecas digitales, proliferan nuevos proyectos sobre bibliotecas digitales y las bibliotecas clásicas comienzan a llevar a cabo proyectos para el manejo de imágenes digitales, documentos y brindar servicios de red (L. Borgman 1999). El rápido desarrollo de las redes de computadoras y las bases de datos, ha influido en el uso y aplicación del término bibliotecas digitales.

Los términos biblioteca virtual, electrónica y digital se suelen utilizar indistintamente como sinónimos, pero no deberían utilizarse como un único concepto. Las **bibliotecas virtuales** son aquellas que solo existen en un espacio informativo virtual (Tramullas Saz 2012). Mientras que (Sánchez Díaz y Vega Valdés 2002) la define como una biblioteca que es creada a partir de documentos digitalizados y sitios que incorpora la realidad virtual, entre sus objetivos y procesos están la selección, adquisición, el procesamiento analítico sintético de la información, la organización de la información y los servicios.

La **biblioteca electrónica** está formada por objetos físicos que necesitarían de medios electrónicos para el acceso a la información contenida en los mismos. Realiza la gestión de todas las funciones de

una biblioteca tradicional de forma electrónica haciendo uso de la tecnología. Puede almacenar diversos contenidos y en diferentes formatos digitales (Tramullas Saz 2012).

Las ***bibliotecas digitales*** se conocen como el lugar donde el usuario puede acceder al universo de conocimientos con mayor rapidez, desde su escritorio o terminal de comunicaciones, por lo que las personas dejan de depender de la biblioteca como edificio. Es una colección de documentos digitalizada, disponible en red con alcance global, por lo que implica una nueva forma de acceder y usar la información. La biblioteca digital es una institución que ha sustituido todos los documentos impresos por sus versiones digitales y realiza todos los procesos básicos de una biblioteca a través de software, para lo que se apoya totalmente en la computación y las redes (Sánchez Díaz y Vega Valdés 2002).

Como concepto a utilizar en la investigación se asume que: una **biblioteca digital** *es un sistema de tratamiento técnico, acceso y transferencia de información digital, estructurado alrededor del ciclo de vida de una colección de documentos digitales, sobre los cuales se ofrecen servicios interactivos de valor añadido para el usuario final* (Tramullas Saz 2012).

Se ha seleccionado este concepto porque se tiene en cuenta el tratamiento de los recursos contenidos en la biblioteca digital. Como ya se ha mencionado anteriormente un recurso en una biblioteca digital puede ser un documento en formato PDF, una imagen, un video u otros. El tratamiento técnico de estos documentos es el proceso de extracción de metadatos, que puede ser realizado manualmente por especialistas en bibliotecología o a través de una aplicación informática.

Los **metadatos** son un componente importante para cualquier sistema de información digital. En (Greenberg, Swauger y Feinstein 2013) se define el término “capital de metadatos”, el cual se refiere a que los metadatos ya sean obtenidos por un proceso manual o automático son considerados un bien público tanto en una biblioteca como en otro entorno ya que apoyan y son un elemento clave en el descubrimiento y acceso a la información. Estos elementos demuestran la importancia de los metadatos y la necesidad de su extracción a partir de documentos digitales y de su manejo en una biblioteca digital.

Las comunidades de las ciencias de la computación, estadística, base de datos, ciencias de la información y bibliotecología han adoptado el concepto de metadatos como “datos sobre datos” (Greenberg 2009). En general los metadatos son objetos que describen algo sobre otro objeto de información (Vásquez Paulus 2015). Los metadatos pueden ser vistos como datos estructurados alrededor de un objeto que admite funciones asociadas al objeto designado, siendo un objeto una entidad, un formulario o modo por el cual los datos pueden ser guardados (Greenberg 2009).

Los **metadatos** surgen como una solución para la recuperación y descripción de la información existente en las bibliotecas digitales (Senso y de la Rosa Piñero 2003). Se definen también como atributos que pueden identificar objetos, almacenados en formato electrónico o no. Estos objetos pueden ser documentos en papel, electrónicos o información de otro tipo (MILLER 1996; HUSBY 1997).

Una definición más acertada y completa de los metadatos, dada por (Senso y de la Rosa Piñero 2003) es: descripción e identificación de los atributos de un recurso, dato u objeto. Un objeto puede ser un recurso bibliográfico (documentos en formato duro o no), registros e inventarios archivísticos, recursos visuales y otros, que incluye información sobre su contexto, contenido y control, para la búsqueda, recuperación, autentificación y evaluación de un recurso.

Se considera el concepto anterior como el más acertado y completo porque resume las características y las funciones de los metadatos. No obstante, se le incorporan los elementos: los metadatos pueden ser obtenidos a partir de un proceso manual o automático y que son de vital importancia para la búsqueda y acceso al conocimiento científico, con el objetivo de dar una definición final de metadatos y de utilizarla en la investigación.

El concepto de metadatos queda definido como: *elementos obtenidos a partir de un proceso manual o automático que describen e identifican los atributos de un recurso, dato u objeto, pudiendo ser un objeto un artículo científico en formato PDF, que incluye información sobre su contexto, contenido y control, para la búsqueda, recuperación, autentificación y evaluación, siendo de vital importancia para la búsqueda y acceso al conocimiento científico.*

Son ejemplos de metadatos (Senso y de la Rosa Piñero 2003):

- El encabezamiento de un fichero multimedia (imagen, video o audio)
- El resumen de un documento
- Catálogo de una base de datos
- Las palabras extraídas de un texto
- Las fichas catalográficas en cualquier formato (ISBD, MARC...)
- Las páginas amarillas

En Internet también se pueden encontrar en multitud de formas:

- Índices de documentos contenidos en una Intranet
- Direcciones IP o DNS
- Directorios X-500

- Encabezamiento de mensajes de correo electrónico
- Descripción de los archivos accesibles vía FTP
- Términos extraídos por los motores de indización/búsqueda

Debido a la relevancia que tienen los metadatos para el proceso de búsqueda y recuperación de la información se hace necesario su extracción de los documentos que los contienen. La **extracción automática de metadatos** *consiste en obtener un conjunto de atributos o elementos necesarios que describan documentos digitales*. Los metadatos extraídos se utilizan para la descripción e identificación de los materiales digitales (Zhang y Zhao 2013). Una vez que este proceso se realiza, estos elementos pueden ser depositados en una base de datos, repositorio o en un lugar donde se almacena información digital con el objetivo de preservarla, generalmente se utiliza el formato RDF (*Resource Description Framework)* y estos metadatos son organizados utilizando tripletas (Lösch, Bloehdorn y Rettinger 2012).

Para llevar a cabo la extracción automática de metadatos desde documentos PDF se pueden aplicar métodos de **Aprendizaje Automático** (ML)[1], sistemas basados en conocimiento y análisis estilístico, tales como, las funciones heurísticas (Sicilia 2014).

El aprendizaje es un fenómeno con varias facetas. El proceso de aprendizaje incluye la adquisición de nuevos conocimientos, el desarrollo de nuevas habilidades a través de la práctica, la organización de nuevos conocimientos en general, representaciones y el descubrimiento de hechos y teorías mediante la observación y la experimentación. El estudio y modelado en computadora de los procesos de aprendizaje en sus múltiples manifestaciones constituye la razón de ser del **Aprendizaje Automático (ML)**. Este campo de estudio de la Inteligencia Artificial tiene como áreas de aplicación fundamentales (Mitchell 1997):

- El desarrollo de sistemas de aprendizaje que lleven a cabo una actividad partiendo de un conjunto de tareas predeterminadas.
- La investigación y la simulación por computadora del proceso de aprendizaje humano.
- La exploración teórica de nuevos métodos de aprendizaje y algoritmos independientemente del dominio de aplicación.

En (Mitchell 2006) se plantea que el ML se centra en la cuestión de cómo crear sistemas de computación que se mejoren así mismos a través de la experiencia adquirida más una estructura inicial

[1] ML, Machine Learning

definida. Además, se encarga de determinar la arquitectura y los algoritmos que pueden ser usados para capturar, almacenar, indexar, recuperar y combinar los datos (Mitchell 2006).

Para (Schapire 2008) el ML trata sobre aprender para hacer mejor una tarea en un futuro basado en experiencias adquiridas y aprendidas en el pasado. También plantea que el objetivo del ML es diseñar algoritmos que realicen el aprendizaje de manera automática sin la intervención o asistencia de una persona.

Los conceptos antes expuestos sobre ML tienen elementos comunes, por lo que se define para la presente investigación que: *El ML es un campo de estudio de la Inteligencia Artificial que tiene como objeto de estudio crear sistemas de computación que sean capaces de aprender para dar mejores soluciones y ejecutar mejor las tareas en un futuro a partir de experiencias adquiridas y aprendidas en el pasado, más una estructura definida en su implementación. El ML se encarga de diseñar y determinar los algoritmos y arquitecturas para que los sistemas de computación sean capaces de aprender sin la intervención o asistencia de una persona.*

1.4. Estado del arte

En el estado del arte de la investigación se realiza un análisis sobre un grupo de herramientas para la extracción automática de metadatos bibliográficos. Como parte del estudio se explican los métodos de aprendizaje automático que utilizan las aplicaciones para llevar a cabo la extracción. También se analizan los lenguajes y estándares empleados para la representación de los metadatos bibliográficos.

1.4.1. Tipos de metadatos

La clasificación de los metadatos por sus tipos o usos todavía no es definitiva, debido al carácter evolutivo que tiene el concepto de metadato según como sean creados y utilizados los mismos (Sicilia 2014). A continuación, se explican tres tipos de metadatos existente en la literatura consultada:

- **Metadatos descriptivos**

Se utilizan para la descripción e identificación de la información contenida en un recurso de información. Contienen atributos físicos (medios, condición de las dimensiones) y atributos bibliográficos (título, autor/ creador, idioma, palabras claves) (Senso y de la Rosa Piñero 2003).

- **Metadatos administrativo**

Se refieren a las características y propiedades del recurso, facilitando la gestión, procesamiento tecnológico y físico de las colecciones digitales tanto a corto como a largo plazo. Incluyen información

sobre la creación y el control de la calidad, la gestión de derechos, el control de acceso, la utilización y las condiciones de preservación (Senso y de la Rosa Piñero 2003).

- **Metadatos estructurales**

Proporcionan información sobre la estructura interna de los recursos electrónicos, como página, sección, capítulo, índice y tabla de contenido, describiendo la relación entre los materiales. Facilitan la navegación y presentación de los recursos y relacionan las diferentes partes que lo componen (Testa y Ceriotto 2011).

De los tres tipos de metadatos analizados, en esta investigación se utilizarán los metadatos descriptivos. Específicamente, de los metadatos descriptivos se usarán sus atributos bibliográficos, ya que estos son los atributos que están contenidos en los artículos científicos.

1.4.2. Fuentes de metadatos

La extracción de metadatos está basada principalmente en el reconocimiento de la estructura de los documentos (Greenberg, Spurgin y Crystal 2006). En (Sicilia 2014) se definen cinco estructuras a partir de las cuales se pueden extraer los metadatos. A continuación, se hace mención de cada una de ellas.

- **Estructura de formato:**

En esta categoría se encuentran documentos que tienen un formato electrónico definido, por ejemplo, en un documento en formato HTML[2] se puede extraer el árbol DOM[3] y encontrar etiquetas como <title> (Sicilia 2014).

- **Estructura visual:**

Los documentos tienen una estructura definida desde un inicio. En esta clasificación se encuentran los documentos en formato PDF, especificando cómo está ubicado el texto en la página. Esto puede ser utilizado para identificar las secciones del documento (Sicilia 2014).

- **Diseño del documento:**

El documento puede estar estructurado siguiendo una disposición específica, por ejemplo, comenzar con el título, luego los autores y finalizar con un número de referencias (Sicilia 2014).

[2] HTML, HyperText Markup Language
[3] DOM, Document Object Model

- **Análisis de citas bibliográficas:**

Los documentos que están relacionados entre sí a través de la vinculación de citas o el análisis de coautoría pueden ser analizados a través de métodos bibliométricos, por lo que es posible tener acceso a varios tipos de información (Sicilia 2014).

- **Estructura Lingüística:**

El documento se puede analizar lingüísticamente, permitiendo deducir el sentido de las partes de las oraciones, o relaciones entre los metadatos. Por ejemplo, las citas en el texto principal pueden estar contenidas dentro de la misma oración, existiendo la probabilidad de que estén relacionadas de alguna manera. La relación puede ser positiva o negativa, dependiendo del texto que lo rodea, por lo que el análisis de la estructura lingüística depende del conocimiento del idioma del documento y posiblemente del conocimiento del dominio. Utilizando un análisis lingüístico se puede intentar extraer las palabras claves y las relaciones entre las citas (Sicilia 2014).

En relación a la estructura visual existen ventajas para HTML, por ejemplo, el trabajo con etiquetas hace más fácil el reconocimiento de estructuras. En formatos como PDF resulta más engorroso ya que especifica símbolos, palabras o ubicaciones de una página y el documento consta de una bolsa de símbolos o palabras en lugares específicos. La estructura del documento se puede inferir de las ubicaciones del símbolo. La desventaja fundamental de estos lenguajes de descripción es que existen múltiples maneras para presentar los textos (Sicilia 2014).

Una vez analizadas las diferentes fuentes desde las cuales se pueden extraer metadatos, se determinó que la fuente más correcta para la investigación es la **estructura visual**, porque de los documentos que están incluidos en esta clasificación se encuentran los documentos en formato PDF, los cuales especifican la organización del texto en la página, permitiendo identificar las secciones o bloques de texto dentro del documento.

1.4.3. Herramientas para la extracción automática de metadatos bibliográficos

Existen varias herramientas dedicadas a la extracción de metadatos bibliográficos a partir de documentos científicos y técnicos en formato PDF, ver **Tabla 2**.

Tabla 2: Herramientas para la extracción automática de metadatos de documentos en formato PDF

Nombre de la herramienta	**Método que utiliza**

Docear´s PDF Inspector*	Análisis del Estilo de la Información (SIA)[4]
GROBID	CRF[5]
Medeley Desktop	SVM, basado en web
ParsCit	CRF
PDFMeat*	Consultas de Google Scholar, pdftotext
SciPlore Xtract*	Análisis de XML

De las herramientas mencionadas en la tabla anterior se seleccionaron aquellas que utilizan técnicas de Aprendizaje Automático. Estas son GROBID, Mendeley y ParsCit. Las herramientas señaladas con un asterisco no serán analizadas como parte del estado del arte de la investigación ya que no utilizan técnicas de ML. Las herramientas seleccionadas para el análisis se explican a continuación.

GROBID[6]

GROBID es un sistema para la extracción y generación automática de metadatos bibliográficos de documentos científicos y técnicos y el reconocimiento de la estructura del documento (López y Romary 2015). Esta herramienta posee una licencia de software libre, desarrollado utilizando el lenguaje de programación Java. Puede ser empleada como una aplicación web o integrada a otros sistemas.

Los metadatos bibliográficos que puede extraer son: los autores, el título, el resumen, las palabras claves, el contenido del artículo, la información sobre la revista en la que fue publicado el artículo y las referencias bibliográficas. Los tipos de documentos utilizados para la extracción de los metadatos en esta herramienta son documentos científicos y técnicos, documentos académicos, manuales técnicos y patentes, siempre en formato PDF.

Uno de sus objetivos es la conversión de los documentos científicos en formato PDF a documentos en formato TEI[7]. Para ello es necesario primeramente reconocer la estructura del documento PDF y luego

[4] SIA, Style Information Analysis
[5] CRF, Conditional Random Fields
[6] GROBID, GeneReration Of BIbliographic Data
[7] TEI, Text Encoding Initiative

extraer los metadatos (López y Romary 2015). En el análisis de las secciones del documento se utiliza el método de aprendizaje automático CRF. GROBID centra su atención en el procesado de las secciones: encabezado (título, resumen), introducción, la sección de títulos, las conclusiones y las referencias bibliográficas, porque en estas secciones los autores introducen los elementos principales y los lectores suelen prestar más atención también a estas partes del documento.

Puede ser utilizado en las bibliotecas digitales como un módulo para el análisis y procesado de documentos de texto. También en las bibliotecas digitales para la obtención de información a partir del procesado de los documentos, para generar y sugerir citas bibliográficas a los usuarios (López 2015).

Mendeley Desktop

Mendeley posee una licencia de software libre. Puede utilizarse a través de un sitio web o una aplicación para PC y dispositivos Apple (IPhone e IPad) para el almacenamiento y manejo de documentos PDF. Permite tener los documentos almacenados en la nube y también compartirlos con otras personas como una red social. La aplicación automáticamente organiza los artículos por categorías (autor, título, revista, fecha y demás) en una base de datos para luego realizar filtrados. Proporciona el manejo de referencias bibliográficas, la selección o creación de estilos de citas textuales y la creación automática de bibliografía (Lo Russo et al. 2013). Permite agregar artículos a la base de datos desde diferentes fuentes, bases de datos online, desde la propia PC o de otras bibliotecas digitales (Lo Russo et al. 2013).

ParsCit

Es una herramienta de código abierto para el análisis de referencias bibliográficas. ParsCit realiza el análisis examinando cada una de las referencias e identificando cada campo que las componen. Los campos extraídos pueden ser utilizados por otros autores. Consta de dos procesos fundamentales para la extracción de las referencias, el preprocesado y el postprocesado (Councill, Lee Giles y Kan 2015).

En el preprocesado, ParsCit utiliza métodos heurísticos para convertir el documento en formato PDF a texto plano, empleando UTF-8 (Councill, Lee Giles y Kan 2015). Luego, en el post-procesado utiliza CRF++, implementación del método de aprendizaje automático CRF, para obtener cada uno de los *tokens* que componen la referencia (Granitzer, Maya y Robert 2015). La herramienta puede ser utilizada tanto como un servicio web o como una aplicación independiente.

1.4.3.1. Comparación entre las herramientas

Con el propósito de conocer qué aplicación tiene un mejor desempeño en el proceso de extracción de metadatos bibliográficos se toma como referencia la comparación hecha por (Lipinski, et al. 2013). Para llevar a cabo la comparación Lipinski seleccionó aleatoriamente una colección de 1153 artículos científicos en PDF, incluyendo sus metadatos, para compararlos con los extraídos por las herramientas estudiadas. Los metadatos seleccionados para el análisis fueron el título, el autor o autores, separando el nombre y los apellidos, el resumen y el año de publicación. Estos metadatos suelen ser los más utilizados en la realización de consultas.

Las herramientas deben cumplir el requisito de permitir la integración con otros proyectos de desarrollo, por ejemplo, una biblioteca digital, a través de una biblioteca de clases o ser una aplicación independiente que permita cargar archivos PDF. A partir de aquí se realizan tres evaluaciones con dos configuraciones de pruebas según el número de artículos que se procesan, cien en la primera y 1153 en la segunda. Los resultados obtenidos para las herramientas seleccionadas se muestran en la tabla siguiente:

Tabla 3: Resultados (A100: Primera evaluación con 100 artículos, B100: Segunda evaluación con 100 artículos, B1153: Segunda evaluación con 1153 artículos), (Lipinski, et al. 2013)

	Título			**Autores**			**Apellidos del autor(es)**		**Resumen**			**Año**	
	A_{100}	**B_{100}**	**B_{1153}**	**A_{100}**	**B_{100}**	**B_{1153}**	**B_{100}**	**B_{1153}**	**A_{100}**	**B_{100}**	**B_{1153}**	**B_{100}**	**B_{1153}**
GROBID	N/A	0.92	0.92	N/A	0.83	0.83	0.90	0.91	N/A	0.75	0.74	0.64	0.69
Mendeley Desktop	N/A	0.84	0.82	N/A	0.72	0.70	0.78	0.77	N/A	N/A	N/A	0.23	0.26
ParsCit	0.59	0.52	0.54	0.47	0.29	0.31	0.36	0.37	0.49	0.31	0.26	0.06	0.07

Los valores representados en la tabla corresponden a la evaluación del desempeño que tuvo cada una de las herramientas en la extracción de los metadatos seleccionados. El valor uno indica que el metadato extraído coincide con los datos referenciados, cero que el metadato fue extraído incorrectamente. De las aplicaciones analizadas GROBID tuvo el mejor desempeño; 0.92 para títulos,

0.83 para los autores, 0.90 para el apellido de los autores, 0.74 para el resumen y 0.69 para el año de publicación.

El desempeño de GROBID indica que los metadatos extraídos tuvieron un mayor nivel de coincidencia con los metadatos que se tomaron como referencia para la comparación. Tiene ventajas sobre las otras herramientas, ya que al trabajar directamente con grandes cantidades de documentos es poca la información que se pierde.

En el artículo (Granitzer, Maya y Robert 2015) se comparan las herramientas Mendeley y ParsCit, obteniendo Mendeley una mejor evaluación. En esta investigación indican que el método SVM es mejor que el método CRF, pero con los resultados obtenidos en la comparación se dice que la implementación de CRF que utiliza GROBID es mejor que el SVM de Mendeley y GROBID tiene un mejor desempeño en la extracción de metadatos que Mendeley.

1.4.4. Métodos de aprendizaje automático para la extracción de metadatos

Las herramientas para el manejo de colecciones de documentos científicos tienen la tarea de extraer los metadatos de cada uno de ellos. Un prerrequisito importante para cumplir esta tarea es el análisis de la estructura del documento. En la actualidad la mayoría de los documentos digitales están en formato PDF, siendo un inconveniente ya que no contienen información sobre su estructura (Klampfl, S., & Kern, R, 2010). Para lograr la extracción de metadatos se utilizan varios métodos, entre las que se pueden mencionar:

- Clasificación, utilizando clasificación bayesiana.
- Coincidencia de patrones, con ejemplos de expresiones regulares.
- Modelos heurísticos.
- Ajustes de modelos, aplicando conocimientos de dominio en la construcción de modelos para la extracción de los metadatos.
- Modelos para obtener la estructura gramatical del documento, utilizando análisis probabilístico, con los modelos: Modelos Ocultos de Markov (HMM)[8], Modelos de Entropía Máxima de Markov[9] y Campos Aleatorios Condicionales (CRF) (Sicilia 2014).

En este trabajo solo se explicarán los modelos HMM, CRF y Máquina de Soporte de Vectores (SVM)[10] ya que siguen un enfoque automático y son los métodos que utilizan las herramientas antes estudiadas.

[8] HMM, Hidden Markov Models
[9] Maximum Entropy Markov Models
[10] SVM, Support Vector Machine

Modelos Ocultos de Markov (HMM)

Es un modelo estadístico que se basa en la modelación de secuencias de datos caracterizadas por la observación de un conjunto de estados. En un documento pueden existir términos ambiguos, o sea, un término puede aparecer en varias partes del texto. Para solucionar este problema los HMM crean una máquina de estados, donde cada estado sería un *token* o palabra en el texto y se asigna una probabilidad de transición de un estado a otro. Con estas probabilidades se examinan todas las posibles evaluaciones de la máquina de estados, y para cada evaluación posible se deja constancia de su probabilidad (Sicilia 2014).

El método HMM es muy útil en el análisis de sistemas donde no se conoce su estructura, permitiendo generar patrones a lo largo del tiempo. Tiene otras aplicaciones en áreas como el procesado de señales y del habla, también se ha utilizado en tareas de NLP (Natural Language Processing), entre las que se encuentran el etiquetado de *part-of-speech*, en la fragmentación de frases y en la extracción de información de documentos (Sicilia 2014).

Un ejemplo práctico de su uso es describir un flujo de texto, observando el modelo del documento científico. Tomando solo el texto del documento quedaría una estructura como una secuencia de datos o tipos, dada por cada campo de texto. Comenzando por el título, el autor(es), dirección de correo, afiliaciones y luego el resumen (Sicilia 2014).

Campos Aleatorios Condicionales (CRF)

Los Modelos Ocultos de Markov están diseñados para trabajar sobre la suposición de que las características del modelo no son independientes, siendo un inconveniente cuando se trabaja con sistemas semiestructurados. Por esta razón son propuestos los *Conditional Random Fields* para el trabajo con modelos donde sus características son independientes (Sicilia 2014).

Son modelos probabilísticos para el etiquetado y segmentación de secuencias de datos. Basado en la definición de una probabilidad condicional $p\ (Y|x)$ sobre una secuencia de *tokens* Y, que es la salida, dada una secuencia de entrada de observación x, en lugar de una distribución conjunta entre el token y la secuencia de observación. Son utilizados para etiquetar una nueva secuencia xs seleccionando la etiqueta ys de manera tal que se logre maximizar la probabilidad condicional $p\ (ys \mid xs)$ (Wallach 2014).

Para el procesado del lenguaje natural se utiliza el etiquetado de partes del contenido, en el cual cada variable ys es una etiqueta de una palabra del texto, en la posición s, y x se divide en xs características. Cada xs contiene información sobre la palabra en la posición s, como su identidad, ortografía, su

pertenencia a un dominio léxico específico e información obtenida de bases de datos semánticas (Sutton y McCallum, 2012).

Utiliza modelos gráficos no dirigidos para las dependencias que existen entre las etiquetas y la estructura que pueden representar en conjunto. Representan la distribución de las variables como un grupo de factores sobre un subconjunto de variables. Usa un enfoque discriminativo en vez de una distribución de probabilidad conjunta sobre la entrada y la salida, para evitar trabajar con una entrada de longitud grande, donde las características representadas en x tengan dependencia entre ellas y esto puede llevar a trabajar con modelos intratables. (Sutton y McCallum, 2012)

CRF ha sido aplicado con éxito en diferentes problemas, entre los que se destacan: procesado de texto, bioinformática y visión artificial. Se han implementado variantes para los modelos CRF. A continuación, se mencionan las más utilizadas (Sutton y McCallum, 2012):

- CRF++
- MALLET
- GRMM
- CRFSuite
- FACTORIE

Máquina de Soporte de Vectores (SVM)

Modelo de aprendizaje supervisado que analiza conjuntos de datos y el reconocimiento de patrones, para la clasificación y análisis de regresión. Es conocido por su buen desempeño en la generalización y en la habilidad de manejar grandes cantidades de datos. Considerando dos clases de un problema de clasificación, siendo $\{(x1, y1), \dots, (xN, yN)\}$ dos clases del conjunto de entrenamiento, xi es el vector de entrenamiento y sus etiquetas yi que pertenecen al intervalo $(-1; +1)$. SVM trata de encontrar el mejor hiperplano que separe las clases del conjunto de entrenamiento. La función de decisión se denomina clasificador y su función núcleo es escrita como $K\ (xa, xb)$ y puede ser un producto interior, Gaussiano, polinomial o cualquier otra función (Hui Han et al. 2003).

1.4.4.1. Análisis de los métodos

Los métodos descritos anteriormente son los más empleados para el proceso de extracción de metadatos bibliográficos (Granitzer et al. 2012). Los métodos CRF y HMM utilizan la técnica de Etiquetado Secuencial, por lo que en (Lafferty, McCallum y Pereira 2001) se dice que CRF y HMM son considerados los más idóneos para el procesamiento de lenguaje natural y la extracción de información

dado por su habilidad de tomar en cuenta clasificaciones anteriores, durante el proceso de análisis del contenido y estructura del documento.

SVM se basa en la técnica de clasificación de texto estándar, o sea, utiliza un vector para la representación del contexto semántico y sintáctico de las características que identifican a un elemento dentro del texto, ignorando las clasificaciones dadas anteriormente y las posteriores (Granitzer et al. 2012). Esta característica de SVM lo pone en desventaja con los métodos CRF y HMM. A pesar de esta desventaja en la investigación realizada en (Granitzer et al. 2012) se determina que SVM tiene un mejor desempeño que CRF para la extracción de metadatos.

1.4.5. Lenguajes y estándares para la representación de metadatos

Para lograr que los metadatos puedan ser manejados y entendidos por una computadora es necesario utilizar lenguajes y estándares para su representación. En el estudio de las principales aproximaciones existentes se realiza un análisis sobre algunos de los lenguajes y estándares existentes.

1.4.5.1. Metalenguajes

Para que los metadatos se materialicen es necesaria la existencia de lenguajes que permitan especificar la sintaxis en la que se definen las estructuras, además de proveer medios para las especificaciones semánticas necesarias (que digan lo que las expresiones sintácticas significan en términos de un modelo). Estos modelos y sintaxis son las que permiten representar las expresiones, hechos, reglas y consultas sobre las descripciones (Vásquez Paulus 2015).

Varias tecnologías aplicadas a la web han extendido las posibilidades y capacidades de los metadatos, aumentando su riqueza en la descripción de los contenidos. Algunas de estas tecnologías son SGML[11] y XML[12] (Senso y de la Rosa Piñero 2003).

SGML es un metalenguaje que permite la creación de diferentes lenguajes de etiquetado a partir de una DTD[13]. Consta de un conjunto de reglas para la descripción de la estructura de un documento de tal forma que pueda ser intercambiado en plataformas computacionales. En SGML un documento está descrito en función de la estructura de las entidades que lo conforman. Estas entidades son organizadas de manera jerárquica determinando la estructura de los elementos del documento (Vásquez Paulus 2015).

[11] SGML, Standard Generalized Markup Language
[12] XML, eXtensible Markup Language
[13] DTD, Document Type Definition

XML, es un formato simple y flexible para la descripción de contenidos digitales. Es considerado extensible porque le permite al usuario definir nuevas etiquetas para la descripción de su contenido. Se define como un lenguaje de marcado porque posibilita la descripción del contenido y la estructura de un texto utilizando un conjunto de etiquetas (Sicilia 2014).

XML es una versión abreviada de SGML, su objetivo se centra en la posibilidad de intercambiar documentos (referenciales o a texto completo) estructurados a través de la Web. Con XML es posible establecer una estructura arbórea con todos los elementos que constituyen un documento para discriminar, rápidamente, los aspectos genéricos de los específicos. Este sistema de representación se ha destacado por ser vital para la generación automática de metadatos en diversos sistemas compatibles, como, por ejemplo, RDF[14] (Senso y de la Rosa Piñero 2003).

1.4.5.2. Estándares

Existen tantos estándares para la representación de metadatos como dominios de información se puedan encontrar en la Web. A continuación, se mencionan los estándares para la representación de metadatos según los dominios de información en los que son utilizados, propuestos por (Sicilia 2014):

1. **Metadatos para el patrimonio cultural**, aplicados a objetos culturales y recursos visuales. Describen recursos de información que provienen de archivos de instituciones. Las bibliotecas digitales están enfocadas en almacenar el patrimonio cultural científico. Algunos ejemplos de estándares en este campo:
   - Estándares de catalogación tradicionales transformados a esquemas MARC21/MARCXML o Esquema para la Modelación de Objetos de Metadatos (MODS)[15].
   - Estándar de metadatos para ayudas como Descripción de Archivos Codificados (EAD)[16].
   - Iniciativa de Codificado de Texto (TEI)[17]
   - Estándares de metadato para la representación de artes visuales, ejemplo, Categorías para la Descripción de Trabajos de Arte (CDWA)[18].
2. **Metadatos para sistemas de información geográficos y geoespaciales**, en este dominio de aplicación se definió un estándar internacional para la representación de este tipo de metadatos, a través de la ISO 19115.

[14] RDF, Resocurce Descritpion Format
[15] MODS, Metadata Object Description Schema
[16] EAD, Encoding Archival Description
[17] TEI, Text Encoding Initiative
[18] CDWA, Categories for the Description of Works of Art

3. **Metadatos para la información del sector público y el gobierno**, donde la información del sector público son datos producidos y acumulados por personal público para ser reusada o integrada a nuevos servicios o productos[19].
4. **Metadatos para sistemas de información de la educación**, abarcan varios estándares para la representación de metadatos con el objetivo de garantizar la interoperabilidad. Metadatos para Objetos de Aprendizaje (LOM)[20], desarrollado por la IEEE, es el principal estándar de representación de metadatos en este dominio.
5. **Metadatos para preservación de recursos digitales**, maneja todos los procesos asociados a la preservación de los recursos digitales y puede ser considerado un estándar que está presente en el resto de los antes mencionados, cada uno de ellos incluye la preservación de la información digital.

De las clasificaciones de estándares antes mencionadas, la 1 y la 4 son las que más están acorde al objeto de estudio de la presente investigación. Los estándares definidos en la clasificación 1 están diseñados para representar los metadatos obtenidos de colecciones de documentos científicos atesorados en una institución o guardados en una biblioteca digital. En la clasificación cuatro están dirigidos a los metadatos obtenidos desde objetos de aprendizajes. De estas clasificaciones se selecciona el número uno para analizar dos de sus estándares, porque incluye la representación de metadatos bibliográficos extraídos de artículos científicos almacenados en una biblioteca digital. A continuación, se explican los estándares TEI y BibTex, el primero para representar todo el contenido de un documento digital y el segundo para las referencias bibliográficas.

Iniciativa de Codificado de Texto (TEI)

Es un estándar para la representación de textos contenidos en documentos digitales. Su objetivo principal es guiar el uso de métodos para la codificación de textos de manera que puedan ser entendidos por una computadora y se enfoca en áreas del conocimiento como humanidades, las ciencias sociales y la lingüística.[21]

Define un conjunto de etiquetas XML y atributos para la codificación de textos contenidos en documentos digitales. Contiene alrededor de 500 elementos para la codificación de textos de cualquier género, año o lenguaje. Los elementos en TEI se dividen en dos categorías, aquellos dedicados a

[19] Ver, http://ec.europa.eu/information_society/policy/psi/index_en.htm
[20] LOM, Learning Objects Metadata
[21] Ver, http://www.tei-c.org/index.xml

representar los metadatos del texto en cuestión y los que se encargan de codificar las características de la estructura del documento, como son las secciones, encabezados, párrafos y otros.[22]

No es monolítico, por lo que puede ser aplicado a varios tipos de textos y está diseñado para ser modular, es decir, los usuarios pueden incorporar conjuntos de características específicas para un tipo de texto. TEI puede ser extensible, al ofrecer la posibilidad a los usuarios de, acorde a sus necesidades, añadir, redefinir o modificar los elementos o los atributos. Es posible también incluir otros lenguajes XML, como MathML o RDF, dentro del documento TEI.

BibTex

Es un estándar para el manejo automático de las referencias y la bibliografía. Establece un diseño genérico para la representación de los datos bibliográficos, ordenándolos alfabéticamente o según el orden en el que aparezcan en el documento. El formato puede ser seleccionado por el usuario y automáticamente las citas y la bibliografía se codifican acorde con el estilo seleccionado.

Utiliza un marcado lógico para separar el estilo de la bibliografía del contenido. Esto permite usar la misma bibliografía con diferentes estilos de citas sin tener que realizar cambios en el documento ni en la bibliografía, aparte del estilo (Fenn 2006).

1.4.5.3. Análisis de los lenguajes y estándares

Los estándares para la representación de los metadatos utilizan los lenguajes o metalenguajes, como también se les conoce, para la descripción de los metadatos que han sido extraídos. XML es un lenguaje que permite representar de manera clara y sencilla la estructura y el contenido del documento, por esta razón es el más utilizado por los diferentes estándares de metadatos. Este lenguaje proporciona reglas para crear nuevos lenguajes de marcado (Johnston 2005), por lo que cada estándar puede definir las etiquetas y estructura que tendrá el documento XML utilizado para describir los metadatos.

La utilización del lenguaje XML permite la interoperabilidad entre varios sistemas. La principal contribución de XML es el intercambio de datos entre sistemas diferentes que utilicen la misma sintaxis (Johnston 2005). Otra de las ventajas de utilizar XML como lenguaje de representación de metadatos es que al permitir el intercambio de metadatos entre sistemas o aplicaciones contribuye a la obtención de información.

[22] Ver, http://www.tei-c.org/Support/Learn/intro

TEI es un estándar que puede ser utilizado para representar cualquier parte del contenido de un documento, así como su estructura. BibTex está dedicado solo para representar las referencias bibliográficas dentro del documento. Esta característica de BibTex lo pone en desventaja con respecto a TEI, por lo que se puede decir que TEI, en este caso, es el mejor estándar para la representación de los metadatos bibliográficos de un artículo científico.

1.5. Conclusiones parciales

La revisión de la bibliografía evidenció la importancia de la extracción de metadatos para la búsqueda, acceso y recuperación de la información principalmente en un ambiente académico o científico. La extracción de metadatos tiene su principal área de aplicación en las bibliotecas digitales para la catalogación de los recursos que almacenan. Con el análisis de los principales conceptos del campo de acción y sus relaciones, se logró profundizar en el problema planteado por la investigación. El estudio de las herramientas para la extracción automática de metadatos evidenció que GROBID es la que mejor realiza el proceso de extracción de metadatos.

CAPÍTULO 2. DESCRIPCIÓN DE LA PROPUESTA

2.1. Introducción

En el presente capítulo se describe la solución que se propone como resultado de este trabajo. Para ello primeramente se hace un análisis sobre su diseño y la arquitectura. Se especifican los principales productos de trabajo generados, así como la propuesta arquitectónica del componente para la extracción automática de metadatos bibliográficos a partir de corpus textuales en formato PDF. Además, se realiza un estudio sobre las herramientas y tecnologías empleadas en el proceso de implementación de la solución y la metodología de desarrollo utilizada.

2.2. Descripción de la propuesta

La solución que se propone como resultado de este trabajo es el desarrollo de un componente para la extracción de metadatos bibliográficos a partir de corpus textuales en formato PDF. Esta aproximación implicaría una reducción en cuanto al costo de tiempo empleado en el proceso de extracción de los metadatos que realizan los especialistas en bibliotecología en una biblioteca.

La propuesta se describe conceptualmente de la siguiente manera. Primeramente, se obtiene un documento en formato PDF o una colección de estos, además de un conjunto de datos necesarios para su manipulación. El o los documentos son analizados en un proceso en el cual se le extraen de forma automática los metadatos bibliográficos. Los metadatos son almacenados en una base de datos relacional y luego son revisados manualmente para comprobar si existen incoherencias con el documento, en caso de existir errores son corregidos y se actualizan en la base de datos.

De la descripción anterior se pueden deducir tres procesos fundamentales:

1. **Introducir datos y documentos en formato PDF**
   Este proceso consiste en introducir los datos relacionados con la procedencia de los documentos y cargar en el sistema un documento o una colección de documentos. Los datos a especificar son el tipo de colección a la que pertenece el documento, o sea si pertenece a una revista o evento científico, además del número y volumen y la edición respectivamente. Los documentos añadidos constituyen la entrada al siguiente proceso que se encargará de su procesamiento.
2. **Procesar documentos**
   En la fase de procesamiento de documentos se lleva a cabo la extracción automática de los metadatos bibliográficos. Este proceso tiene como entrada los documentos obtenidos en la fase inicial. Los documentos son procesados utilizando la herramienta GROBID. GROBID genera un

documento XML que contiene los metadatos correspondientes a un documento. El archivo XML es analizado utilizando un *parser* que se encarga de obtener los metadatos. Finalmente, los metadatos son almacenados en una base de datos relacional para su posterior revisión.

3. **Catalogar metadatos**

   El proceso de catalogación es donde el usuario debe revisar si los metadatos extraídos están en correspondencia con el documento procesado. El usuario selecciona el documento y a continuación se muestran los metadatos correspondientes al documento. Los metadatos pueden ser editados si están incorrectos y se actualizan directamente en la base de datos.

2.2.1. Diagramas de procesos y flujo de datos

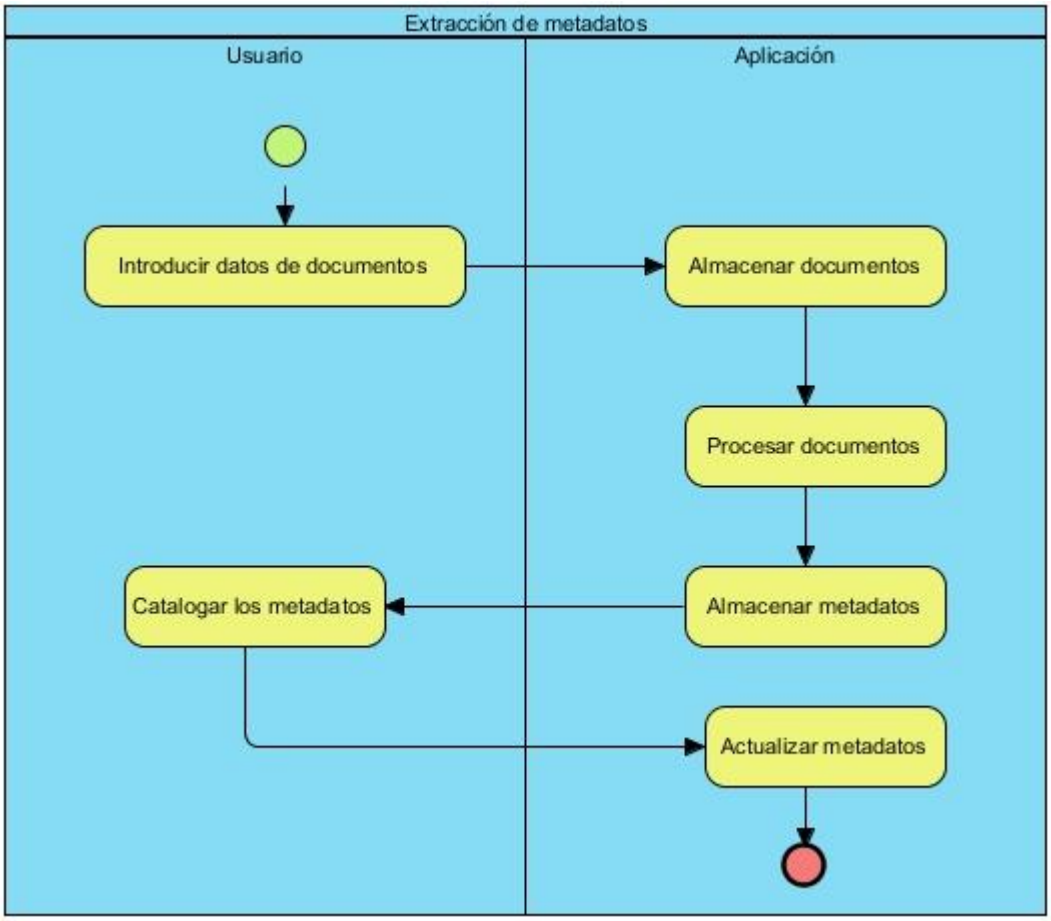


Figura 2: Diagrama del proceso de extracción de metadatos bibliográficos

En la **Figura 2** se describe el funcionamiento de la aplicación a través de los diferentes procesos que la componen. El flujo comienza con el proceso *Introducir datos de documentos* y culmina con el subproceso *Actualizar metadatos*. Son representados los subprocesos *Almacenar documentos*, *Almacenar metadatos* y *Actualizar metadatos*. Estos subprocesos son agrupados

por los tres procesos principales explicados anteriormente. El diagrama además especifica los actores que intervienen en cada proceso.

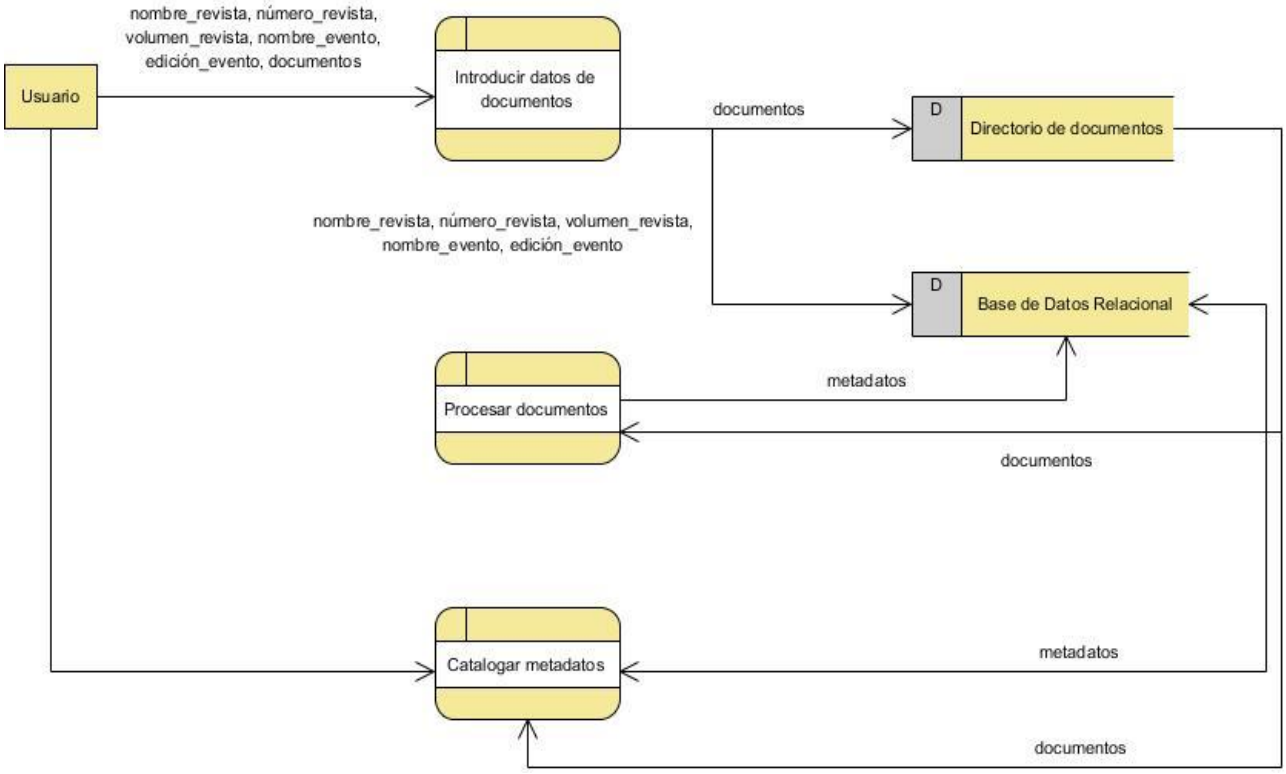


Figura 3: Flujo de datos del Componente para la Extracción de Metadatos Bibliográficos

El flujo de datos en el componente se muestra en la **Figura 3**. El comportamiento de los datos es representado a través de flechas que indican hacia dónde van los datos. En cada una se especifican los datos que son manejados en el componente. En el diagrama están presentes los tres procesos fundamentales del componente y la base de datos relacional y el directorio de documentos que se va a utilizar.

2.3. Metodología de desarrollo de software

El desarrollo de todo software debe estar guiado por una metodología de desarrollo para así lograr que este tenga la calidad requerida. Existen dos grupos de metodologías: ágiles y tradicionales. No existe una metodología universal para cada tipo de proyecto. Se define una metodología según las características del equipo de desarrollo, el dominio de aplicación, el tipo de contrato, la complejidad y la envergadura del proyecto.

Se decide optar por el uso de una metodología ágil, lo cual se justifica a través de los siguientes elementos:

- La necesidad de desarrollar la propuesta de solución en un período corto de tiempo, con un equipo de desarrollo pequeño.
- Garantizar la flexibilidad necesaria en cuanto a la variación de los requisitos.
- El cliente forma parte del equipo de desarrollo, lo que permite reducir la generación de documentos y artefactos.

Se selecciona para guiar el desarrollo de la solución propuesta la metodología AUP-UCI que responde a un enfoque ágil. Esta es una variación de la metodología AUP que apoya al ciclo de vida definido para la actividad productiva de la UCI. AUP-UCI consta de tres fases: Inicio, Ejecución y Cierre y propone siete disciplinas, entre las que se encuentra Requisitos.

La disciplina Requisitos propone cuatro escenarios para la modelación del sistema en los proyectos. El más idóneo para la modelación de los requisitos de la solución propuesta es el escenario cuatro, Historias de Usuarios. Se selecciona este porque existe una lógica de negocio bien definida, el cliente siempre acompaña al equipo de desarrollo para convenir los detalles de los requisitos, poder implementarlos, probarlos y validarlos, además que el proyecto no es muy extenso.

2.4. Entorno de desarrollo

A lo largo de la investigación se utilizan un grupo de herramientas y tecnologías. A continuación, se realiza una breve descripción de cada una de ellas.

Sistema Gestor de Base de Datos

Un Sistema Gestor de Bases de Datos (SGBD) es una colección de programas cuyo objetivo es servir de interfaz entre la base de datos, el usuario y las aplicaciones. Posee un lenguaje para la manipulación y consulta de los datos. Permite definir los datos a distintos niveles de abstracción.

PostgreSQL[23] es un sistema de código abierto para la gestión de bases de datos relacionales. Software multiplataforma, puede correr en varios sistemas, Linux, UNIX (AIX, BSD, HP-UX, SGI IRIX, Mac OS X, Solaris, Tru64) y Windows. Posee interfaces nativas para la programación en C/C++, Java, .Net, Perl, Python, Ruby, Tcl, ODBC y otros.

[23] Ver, https://www.postgresql.org/

El uso de este SGBD ofrece ventajas sobre otros sistemas de bases de datos. Además de ser software libre, tiene una amplia comunidad que contribuye en el soporte de la herramienta, alta estabilidad y potencia, facilidad de administración e implementación de estándares. En el desarrollo de la solución se utiliza la versión PostgreSQL 9.4.

Marco de trabajo

Grails[24] es un marco de trabajo para el desarrollo web de gran alcance, para la plataforma Java destinada a multiplicar la productividad de los desarrolladores gracias a un convenio-sobre-configuración, los parámetros por defecto y las API. Se integra sin problemas con la Máquina Virtual de Java (JVM)[25]. Proporciona características de gran alcance, incluyendo Modelo de Objeto Relacional (ORM)[26], tiempo de ejecución y meta-programación en tiempo de compilación y programación asíncrona. Para el desarrollo de la propuesta de solución se utiliza la versión 2.5.3.

Lenguaje de Programación

Se utiliza el lenguaje de programación **Groovy**[27] en su versión 2.5. Es un lenguaje dinámico y puede ser tipado o no tipado, con capacidades de compilación estáticas. Está orientado a mejorar la productividad del desarrollador gracias a su sintaxis concisa, familiar y fácil de aprender. Se puede integrar fácilmente a cualquier aplicación Java.

Está diseñado para el desarrollo de aplicaciones web, aplicaciones interactivas, marcos de prueba, construir herramientas, análisis de código y para el diseño de interfaces de usuario. Entre sus características más relevantes se encuentra el uso de *closures*, programación funcional, meta-programación, inferencia de tipos y compilación estática.

Entorno de desarrollo Integrado

Un Entorno de Desarrollo Integrado, por sus siglas en inglés (IDE) es un programa compuesto por una serie de herramientas utilizadas por programadores para desarrollar aplicaciones.

IntelliJ Idea[28] es un entorno de desarrollo integrado, por sus siglas en inglés (IDE). Consiste en un software cuyo principal objetivo es el desarrollo de otro software. Permite el desarrollo de aplicaciones web y para empresas, incluye además programación en lenguajes de la máquina virtual de Java, JVM por sus siglas en inglés, Java, Scala, Groovy y Kotlin. Proporciona herramientas integradas para el

[24] Ver, https://grails.org/
[25] JVM, Java Virtual Machine
[26] ORM, Object Relational Model
[27] Ver, http://www.groovy-lang.org/
[28] Ver, https://www.jetbrains.com/idea/

trabajo con sistemas de control de versiones y para el manejo de bases de datos a través del lenguaje SQL. En este trabajo se utiliza la versión IDE IntelliJ Idea 14.0.1, la cual es definida por el grupo de investigación de Web Semántica.

2.5. Diseño de la propuesta de solución

La solución que se propone como resultado de la investigación es el desarrollo de un componente para la extracción automática de metadatos bibliográficos desde corpus textuales en formato PDF. Antes de comenzar el desarrollo de la solución propuesta es necesario definir y validar los requisitos funcionales y no funcionales, para ello se utilizan un conjunto de técnicas. Una vez definidos los requisitos se procede a su descripción a partir de las historias de usuario que propone el escenario número cuatro de la metodología AUP-UCI, estimando el tiempo y el esfuerzo que se necesita para implementar cada una. Se describe la arquitectura que sigue el componente para la extracción de metadatos bibliográficos.

2.5.1. Requisitos

La especificación de requisitos del software es una descripción completa del comportamiento del sistema software a desarrollar. Incluye la descripción de todas las interacciones que se prevén que los usuarios tendrán con el software. También contiene requisitos no funcionales, que imponen restricciones de diseño o funcionamiento del sistema software (Pytel et al. 2011).

2.5.1.1. Técnicas para la captura de requisitos

La captura de requisitos es la actividad donde el analista del equipo de desarrollo trabaja junto al cliente para describir el problema que el sistema debe resolver, los diferentes servicios que el sistema debe proporcionar, las restricciones que se pueden presentar y otros. Esta actividad debe ser muy efectiva ya que de su éxito dependerá que se satisfagan las necesidades del cliente (Chave 2007).

Existen varias técnicas para la captura de requisitos, de ellas solo se seleccionó para ser aplicada la entrevista:

- **Entrevista**: esta técnica se aplicó con el objetivo de tener un mejor entendimiento del problema y de las necesidades del cliente. A partir de estas se definen, los requisitos funcionales que debe cumplir la propuesta de solución. La entrevista se realizó a partir de la selección de un conjunto de preguntas que responden a lo anterior y fueron realizadas al MSc. Yusniel Hidalgo Delgado, jefe del grupo de investigación de Web Semántica, que a su vez es el cliente y tiene más conocimientos sobre el objeto de estudio y campo de acción de la investigación.

La entrevista realizada brindó información que es fundamental para el desarrollo de la propuesta de solución. A partir de ella se definieron los requisitos funcionales de la solución, los cuales constituyen la base para lograr una correcta implementación y así cumplir con las necesidades y expectativas del cliente.

2.5.1.2. Definición de los requisitos funcionales

Los requisitos funcionales de un sistema son aquellos que especifican lo que este debe hacer. Describen las transformaciones que el sistema realiza sobre las entradas para producir las salidas. La tabla siguiente muestra los requisitos funcionales de la propuesta de solución:

Tabla 4: Requisitos funcionales

No.	Requisito	Prioridad	Complejidad
1	Introducir datos y documentos	Alta	Media
2	Procesar documentos	Alta	Alta
3	Catalogar metadatos	Alta	Media

2.5.2. Historias de Usuario

Las historias de usuario (HU) describen los requisitos del sistema en un lenguaje fácil de entender para el cliente y a su vez le indiquen, con claridad, qué hacer a los desarrolladores. El cliente establece la prioridad de cada HU y los programadores realizan una estimación del esfuerzo necesario de cada una de ellas. La prioridad de cada HU es la clasificación más importante, ya que está directamente relacionada con los intereses del cliente. Las HU pueden ser divididas en dos o más historias, si el programador entiende que no es lo completamente sencilla como para implementarla como una funcionalidad atómica.

Para la solución que se propone en la investigación se definen cuatro HU, las cuales son explicadas en el acápite ***Plan de iteraciones***. Estas HU son definidas a partir de los tres procesos descritos al inicio del capítulo. El requisito *Procesar documentos* se desglosa en otros dos, ya que tiene una complejidad de implementación alta. Los requisitos son: *Procesar documentos* y *Procesar archivos XML* y son descritos por las HU con el mismo nombre.

2.5.2.1. Estimación de esfuerzo por Historias de Usuario

Según la prioridad asignada por el cliente a cada HU y teniendo en cuenta la complejidad determinada por el programador, se realiza la estimación de cada una de las HU identificadas. Los resultados de la estimación se muestran en la siguiente tabla. La unidad de estimación es el punto. Un punto equivale a una semana ideal de programación.

Tabla 5: Estimación por HU

Historias de Usuario	Puntos de estimación
Introducir datos y documentos	2
Procesar de documentos	3
Procesar archivos XML	3
Catalogar metadatos	2

2.5.2.2. Plan de iteraciones

Una vez descritas las HU y su estimación por el cliente para su posterior implementación por los desarrolladores involucrados, se procede a la planificación de la entrega del sistema, agrupando todas las HU que serán desarrolladas por cada iteración. A partir de lo planteado se decide desarrollar el sistema en dos iteraciones, las cuales se especifican a continuación. Cada iteración está asociada a los procesos fundamentales del componente a desarrollar.

Iteración 1

En esta iteración se implementan las HU *Introducir datos de documentos*, *Extraer metadatos bibliográficos* y *Procesar archivos XML* descritas en las **Tabla 6**, **Tabla 7** y **Tabla 8**. La primera permite obtener los datos que identifican a un documento como la revista o el evento al cual pertenece, así como el documento en sí o una colección de estos. Las restantes HU corresponden al proceso *Procesar documentos*. Estos requisitos en conjunto se encargan de extraer y almacenar los metadatos bibliográficos de los documentos.

Tabla 6: HU Introducir documentos

Número: 1	**Nombre del requisito:** Introducir datos y documentos.
Programador: Leduan Flores Riera	**Iteración Asignada:** *1*
Prioridad: ***Alta***	**Tiempo Estimado:** *2*
Riesgo en Desarrollo: ***Alto***	**Tiempo Real:** *1*

Descripción: El usuario introduce los datos sobre el documento y carga de un archivo externo el o los documentos en formato PDF que desea procesar. Los datos son la revista, con el volumen y número, o el evento y su número de edición al que pertenece. Los documentos son almacenados en un directorio para su posterior utilización.

Observaciones:

Prototipo de interfaz:

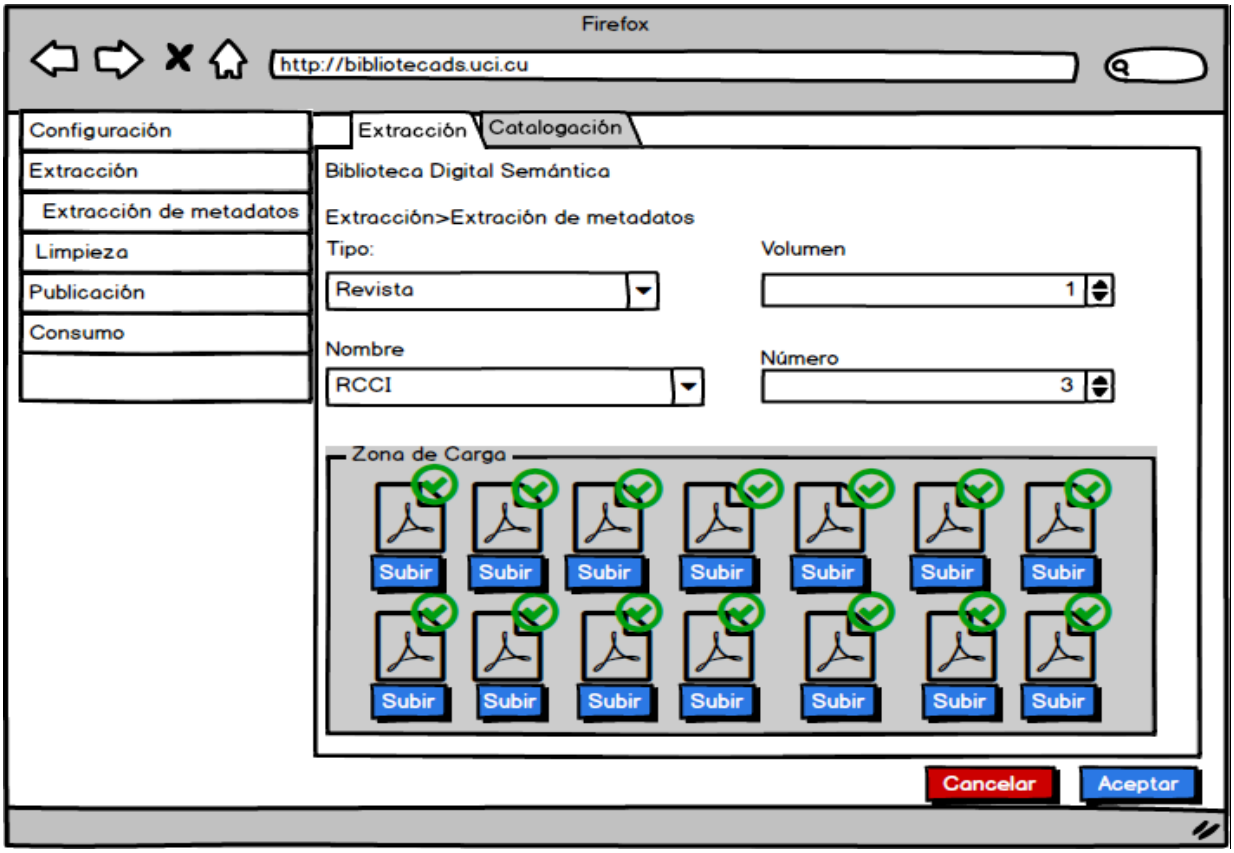


Tabla 7: HU Procesar documentos

Número: 2	**Nombre del requisito:** Procesar documentos.
Programador: Leduan Flores Riera	**Iteración Asignada:** *1*
Prioridad: ***Alta***	**Tiempo Estimado:** *3*
Riesgo en Desarrollo: ***Alto***	**Tiempo Real:** *4*
Descripción: Los documentos introducidos por el usuario son almacenados en un directorio de documentos. Luego los documentos son procesados con el fin de extraer sus metadatos bibliográficos. Una vez extraídos los metadatos estos son representados utilizando el lenguaje XML, estos archivos son procesados posteriormente.	
Observaciones:	
Prototipo de interfaz: No necesita un prototipo de interfaz ya que el proceso es transparente al usuario. Los datos de salida son mostrados en el prototipo de interfaz correspondiente al requisito Catalogar metadatos.	

Tabla 8: HU Procesar archivos XML

Número: 3	**Nombre del requisito:** Procesar archivos XML.
Programador: Leduan Flores Riera	**Iteración Asignada:** *1*
Prioridad: ***Alta***	**Tiempo Estimado:** *3*
Riesgo en Desarrollo: ***Alto***	**Tiempo Real:** *4*
Descripción: Los metadatos son presentados utilizando archivos XML, siendo un inconveniente para su almacenamiento. Estos archivos se analizan para extraer su contenido. Los metadatos son almacenados en una base de datos relacional.	

Observaciones:
Prototipo de interfaz: No necesita un prototipo de interfaz ya que el proceso es transparente al usuario. Los datos de salida son mostrados en el prototipo de interfaz correspondiente al requisito Catalogar metadatos.

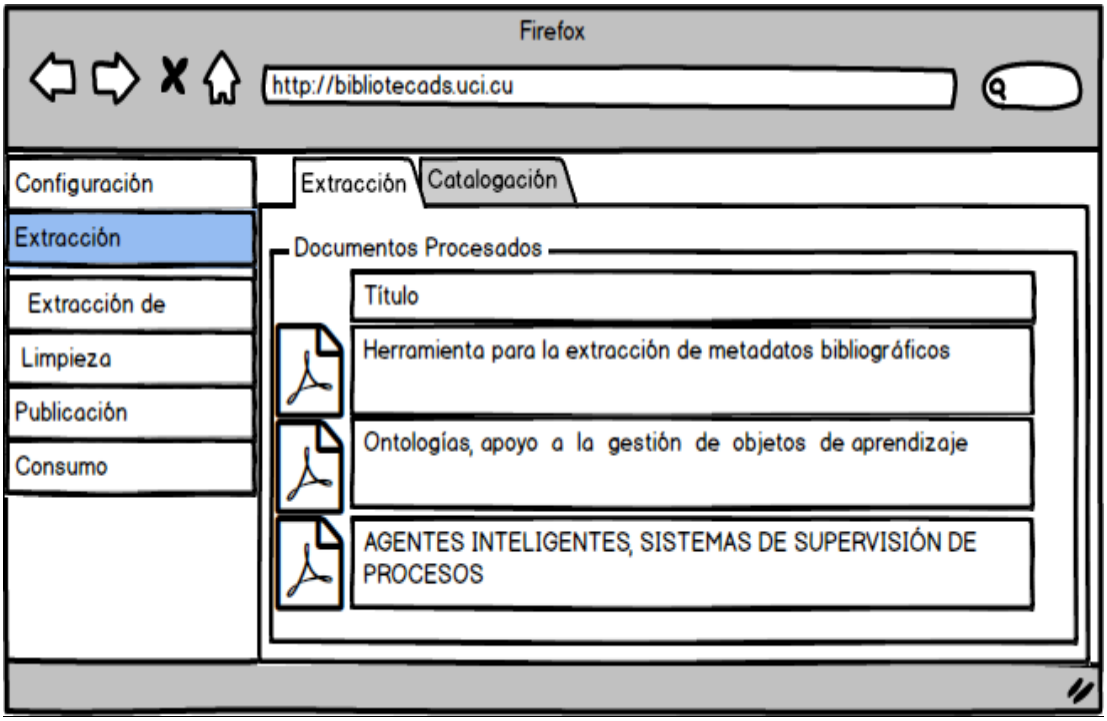
Iteración 2

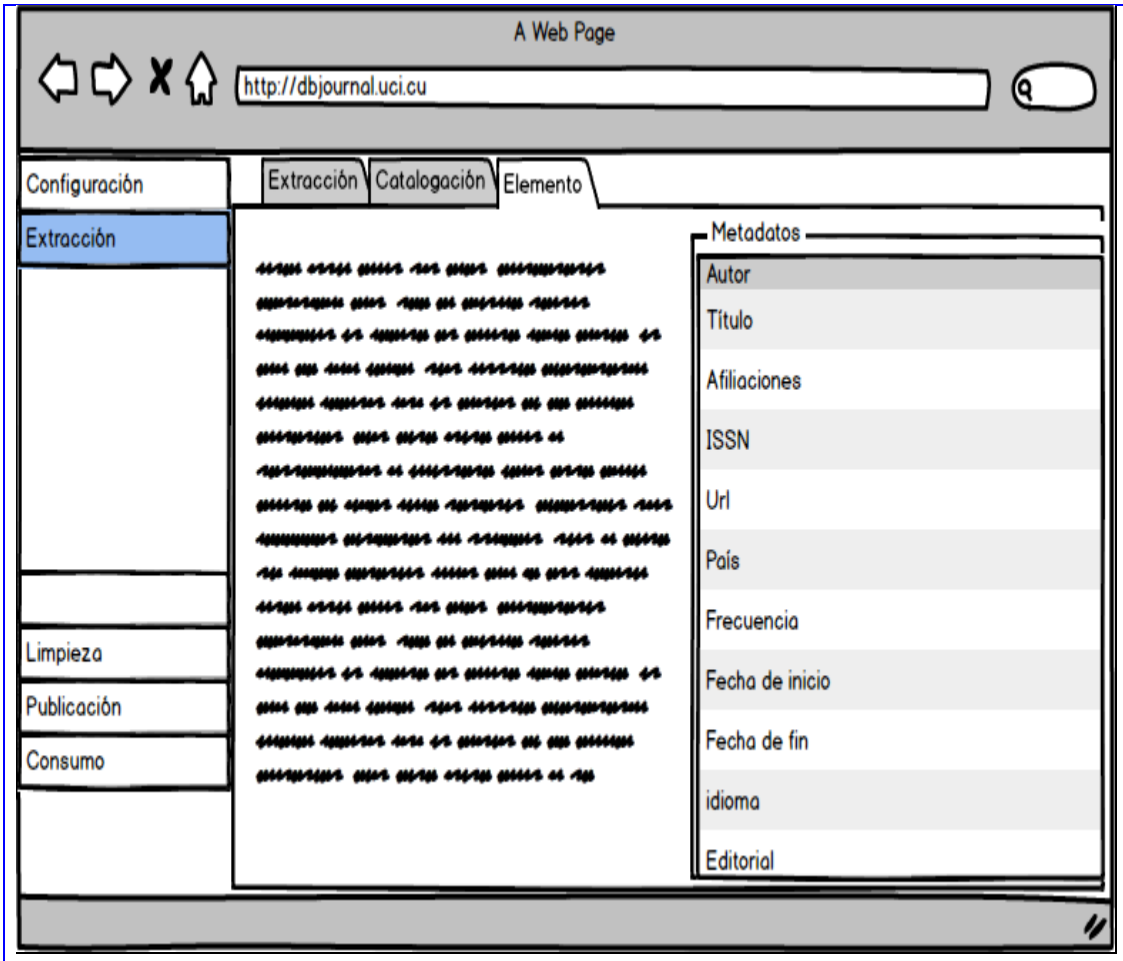
En la segunda y última iteración se implementa la HU *Catalogación de metadatos bibliográficos*, descrita en la **Tabla 9**. Con esta HU se pretende la revisión de los metadatos por parte del usuario.

Tabla 9: HU Catalogación de metadatos bibliográficos

Número: 4	**Nombre del requisito:** Catalogar metadatos.
Programador: Leduan Flores Riera	**Iteración Asignada:** *2*
Prioridad: ***Alta***	**Tiempo Estimado:** *2*
Riesgo en Desarrollo: ***Alto***	**Tiempo Real:** *1*
Descripción: Una vez extraídos los metadatos el usuario revisa manualmente si los metadatos extraídos por la herramienta están correctos, comparándolos directamente con el documento PDF correspondiente. El documento debe ser seleccionado previamente sin tener que esperar a que se terminen de procesar todos los documentos.	
Observaciones:	

Prototipo de interfaz:





2.5.3. Definición de los requisitos no funcionales

En este acápite se describen los requisitos no funcionales. Los requisitos no funcionales son restricciones que debe cumplir la aplicación. Consisten en restricciones impuestas por el entorno y las tecnologías. Pueden ser especificaciones sobre tiempo de respuesta o volumen de información tratado por una unidad de tiempo, requisitos en cuanto a interfaces, extensibilidad y facilidad de mantenimiento. Para la solución se proponen los siguientes requisitos no funcionales:

Apariencia o interfaz externa

Las interfaces de usuario deben ser de color rojo, negro, azul o blanco. Los botones son azules, rojos o verdes. Los textos deben ser escritos usando las fuentes openSans, sans-serif, en colores blanco y negro. El tamaño de la letra debe ser 14px.

Hardware

PC servidora: El hardware donde será instalada la solución propuesta debe ser de altas prestaciones. El servidor debe contar con un microprocesador Intel Core i5 de tercera generación, una memoria RAM de 4Gb, capacidad de disco duro de un 1Tb y una tarjeta de red con velocidad de 100 Mb/s.

PC cliente: debe tener como requisito de hardware una tarjeta de red con velocidad de 100 Mb/s para establecer las conexiones con el servidor.

Software

PC cliente: debe tener instalado un navegador web, ejemplo Firefox en su versión 44 o superior.

PC servidora: servidor web Tomcat Server, versión 7.9, debe tener el gestor de base de datos PostgreSQL v9.4 o superior. Además, se debe instalar en el servidor la versión siete de la máquina virtual de Java.

Usabilidad

Presenta un menú compuesto por elementos que representan de forma gráfica y textual, cada opción que permite el sistema.

2.5.4. Validación de los requisitos funcionales

La validación tiene el objetivo de verificar si los requisitos descritos son una representación aceptada de la propuesta de solución. Una vez culminada la validación es posible conocer el nivel de conformidad que tiene el cliente con los requisitos. Con este fin se utilizaron las siguientes técnicas:

2.5.4.1. Validación por prototipo de interfaz de usuario

Se diseñaron prototipos de interfaz de usuario que ilustran el resultado final una vez implementada la propuesta de solución según los requisitos funcionales obtenidos. Los usuarios finales y el cliente son los encargados, a partir de la experimentación con estos prototipos y un conjunto de datos seleccionados, determinar el nivel de conformidad que tienen con los requisitos, o sea, si cumplen o no sus necesidades y expectativas. Los resultados obtenidos fueron satisfactorios, los requisitos funcionales en su totalidad cumplen con las necesidades del cliente. En la **Figura 4** se muestra uno de los prototipos de interfaz utilizado para la validación de los requisitos, correspondiente al requisito *Introducir datos y documentos*

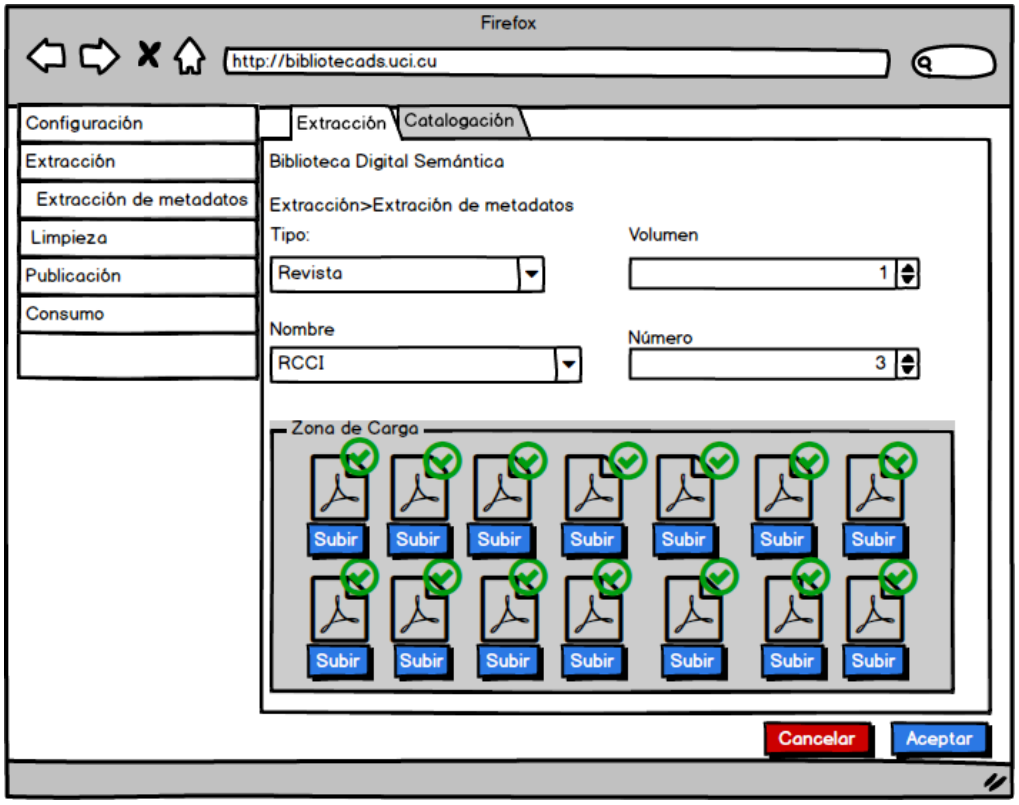


Figura 4: Prototipo para introducir el o los documentos en formato PDF, así como el tipo de colección a la que pertenecen y el nombre de la colección

2.5.4.2. Validación de requisitos por casos de prueba

Se diseñan un conjunto de casos de prueba para comprobar si los requisitos identificados se corresponden con las necesidades del cliente. La ejecución de las pruebas permite encontrar no conformidades que existan en la implementación de los requisitos. Los casos de pruebas creados, su utilización y las no conformidades obtenidas son explicados en el **Capítulo 3**, en el acápite correspondiente a las pruebas de caja negra.

2.5.5. Propuesta de arquitectura

La propuesta de solución sigue un estilo arquitectónico de flujo de datos. Es aplicado en cada proceso desarrollado en el componente. Los datos de entrada de un proceso son transformados en datos de salida que serán la entrada al próximo proceso para su manipulación. El patrón arquitectónico utilizado es tuberías y filtros.

Con el patrón tuberías y filtros cada etapa de procesamiento se encapsula en un filtro. Cada filtro se encarga de procesar los datos que recibe como entrada para transformarlos en datos de salida. Los

datos son transmitidos a través de tubos a los filtros adyacentes para así continuar con el flujo de procesamiento de los datos (Pressman 2006).

En la **Figura 5** se muestra el diseño arquitectónico que sigue la propuesta de solución. Como datos de entrada a la arquitectura se tienen un documento o varios de ellos en formato PDF, además de un grupo de datos que indican si el documento pertenece a una revista o evento científico específico. Estos datos de entrada son manejados por el filtro entrada de datos y documentos. Los documentos se guardan en un repositorio de documentos y los datos sobre la revista o el evento son obtenidos a partir del repositorio de metadatos.

Luego de almacenados los documentos PDF, estos pasan a ser procesados por el filtro procesamiento de documentos. Este filtro se encarga de extraer los metadatos de cada uno de los documentos. Para la extracción de los metadatos este filtro utiliza la herramienta Grobid. Esta herramienta es integrada a la propuesta de solución y como resultado genera un archivo XML donde están cada uno de los metadatos de un documento, los cuales pueden ser: el título, cada uno de los autores, sus afiliaciones o instituciones a las que pertenece cada autor y otros. Los archivos XML son analizados para obtener los metadatos los cuales son almacenados en el repositorio de metadatos.

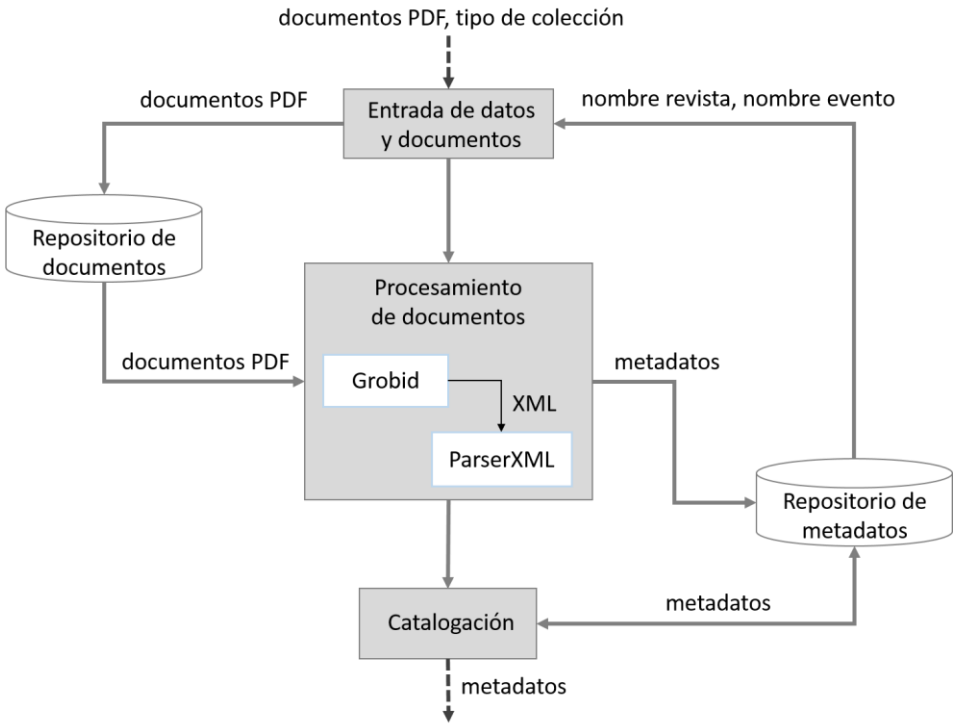


Figura 5: Arquitectura de la propuesta de solución

Los metadatos extraídos en el filtro procesamiento de documentos no siempre son correctos, pueden no ser totalmente extraídos o ser intercambiados unos por otros. Teniendo en cuenta lo anterior se propone el filtro catalogación de metadatos. Este filtro muestra los metadatos al usuario para que a partir del documento correspondiente los corrija. Una vez corregidos son actualizados en el Repositorio de metadatos.

2.5.6. Modelo de datos

En la **Figura 6** se presenta el modelo de datos a utilizar en la propuesta de solución. De las clases representadas en el modelo se utilizan principalmente en la solución: *Revista*, *Evento*, *Volumen*, *Número*, *Edición*, *Persona*, *Documento*, *Institución*, *Documento Persona* y *Palabras clave*. Estas clases son representadas en la propuesta de solución a través de las clases de dominio y dan nombre a las tablas de la base de dato relacional en donde se guardan los metadatos una vez extraídos. Además de estas, se crea la clase de dominio *MetadataExtraction*, que se encarga de almacenar temporalmente los metadatos.

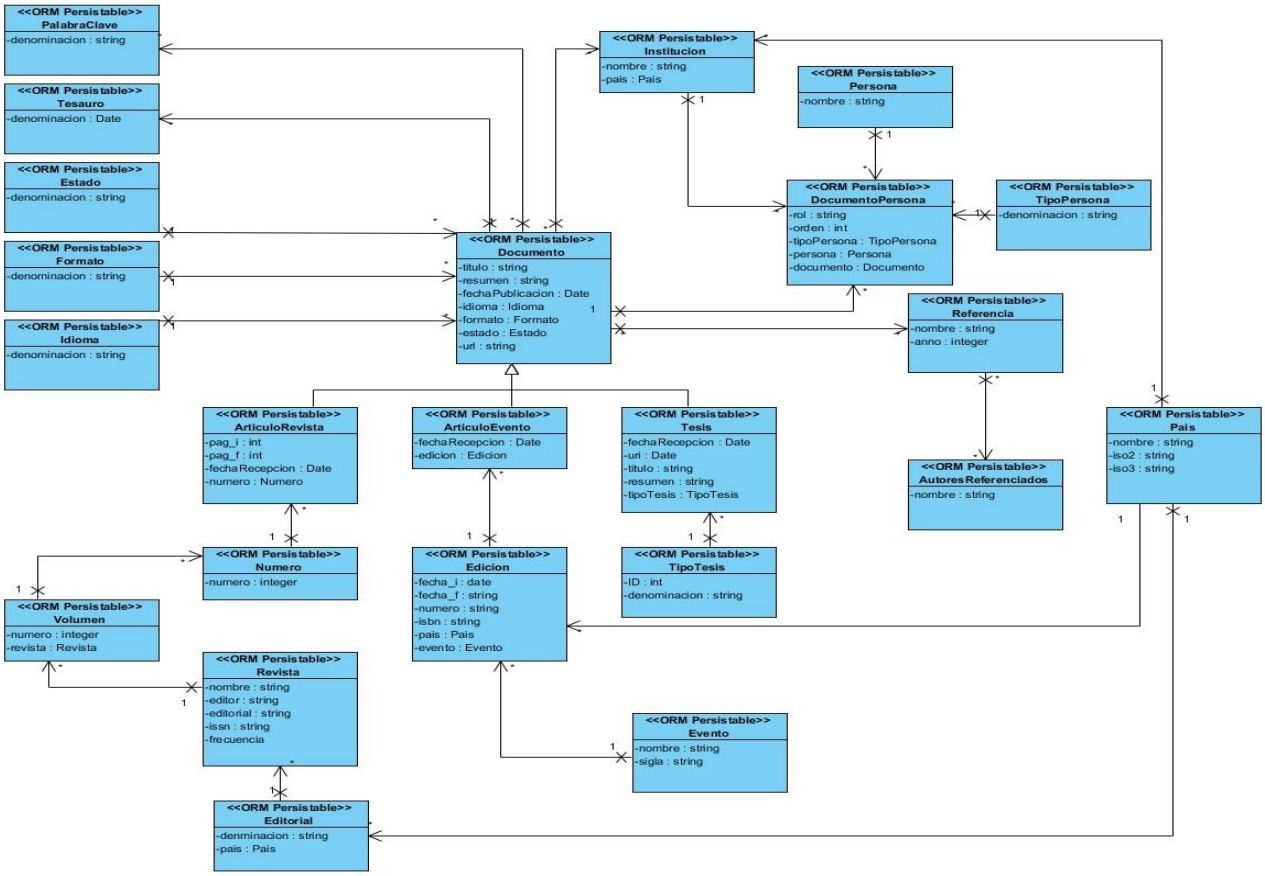


Figura 6: Modelo de datos de la propuesta de solución

Los metadatos extraídos son almacenados en la base de datos en la tabla *MetadataExtraction* hasta que sean actualizados por el usuario en el proceso de catalogación. Terminado el proceso de catalogación los metadatos son guardados en las tablas mencionadas anteriormente y son eliminados de esta.

2.5.7. Estándares de código

Un estándar de código es una forma de normalizar la programación de manera que cualquier persona con acceso al código pueda comprenderlo. Define la escritura y organización del código fuente, cómo deben ser declaradas las variables, las clases y los comentarios. En este acápite se describen los estándares de codificación a utilizar en el proceso de implementación de la propuesta de solución.

2.5.7.1. Nomenclatura de las clases

Se utilizan nombres simples y descriptivos, comienzan con una letra mayúscula y el resto en minúscula, en caso de que tenga más de una palabra se utiliza la notación *UpperCase*. Esta notación indica iniciar con letra mayúscula cada una de las palabras que componen el nombre de la clase.

Ejemplo: ArticuloEvento

2.5.7.2. Nomenclatura según el tipo de clases

La nomenclatura para las clases controladoras y de dominio sigue las pautas descritas anteriormente. A continuación, se explican otros elementos de su nomenclatura.

Clases controladoras: Las clases que se encuentran dentro de la carpeta *controllers* y además en el nombre se incluye la palabra *Controller* para indicar el tipo.

Ejemplo*: ExtractionController*

Clases de dominio: Las clases se guardan en la carpeta *domain* y estas dan nombre a las tablas de la base de datos.

Ejemplo: Revista

2.5.7.3. Nomenclatura de funcionalidades y atributos

El nombre a emplear para las funciones y los atributos se escriben con la inicial del identificador en minúscula, en caso de que sea un nombre compuesto se empleará notación *CamelCasing*. Se utilizan verbos para nombrar a las funcionalidades, pero no en el caso de los atributos.

Ejemplo de función: *getTitle()*

Ejemplo de atributo: *title*

2.5.7.4. Normas de comentarios

Se utilizan comentarios para brindar información adicional al desarrollador acerca de la implementación de cada clase, método, propiedad o constante creada. Esto permite un mayor entendimiento del código para otros programadores y posibilita el mantenimiento y evolución a lo largo del tiempo.

2.6. Conclusiones parciales

La definición de los requisitos funcionales permitió tener una mejor comprensión de las necesidades del cliente, así como determinar cuáles eran las principales funcionalidades que eran necesarias implementar para dar cumplimiento a cada uno de ellos. El patrón arquitectónico seleccionado para representar la arquitectura de la propuesta de solución ilustra cómo se realiza la conversión de la información entrante en datos de salida a través de un flujo de datos. Se integra GROBID a la propuesta de solución como parte del proceso de extracción de metadatos bibliográficos. Como parte de la propuesta de solución se implementó un *parser* para el análisis de los archivos XML obtenidos como resultados del procesamiento de los documentos PDF por la herramienta GROBID.

CAPÍTULO 3. VALIDACIÓN DE LA PROPUESTA

3.1. Introducción

En este capítulo se realiza la validación de la solución propuesta en el capítulo anterior, en conjunto con la planificación de las pruebas de aceptación e internas. Para cada HU se realizan las pruebas de aceptación que describe la metodología AUP-UCI. Los errores detectados por los casos de prueba fueron mitigados tras un proceso iterativo. Se presenta un caso de estudio realizado para evaluar el costo en tiempo que demora la propuesta de solución en realizar la extracción de metadatos bibliográficos y un grupo de participantes.

3.2. Pruebas de software

Las pruebas de software aplicadas a un software independientemente de la metodología de desarrollo utilizada durante todo el proceso de construcción del mismo, permiten aumentar la calidad del sistema eliminando los errores y las no conformidades encontradas durante esta fase. En el caso de la metodología AUP-UCI se proponen tres tipos de pruebas: Internas, Aceptación y Liberación. Para la validación de la propuesta de solución se utilizan las pruebas de Aceptación e Internas, además de las pruebas de Integración para validar la integración de la herramienta Grobid con el componente.

3.2.1. Pruebas Internas

En esta disciplina se verifica el resultado de la implementación y si el producto desarrollado tiene la calidad necesaria para ser utilizado por los usuarios en un ambiente real. Existen distintos tipos de pruebas internas que se le realizan al código para encontrar errores durante su ejecución. En este caso se utilizan las pruebas unitarias y las pruebas de caja negra, las cuales son descritas a continuación:

3.2.1.1. Pruebas unitarias

Las pruebas unitarias son un método para la prueba modular del comportamiento funcional de un programa, o sea, permiten conocer si un módulo del programa está listo y correctamente terminado. Un programa se descompone en unidades, donde cada unidad es una colección de funciones, y las unidades se prueban de forma independiente (Koushik, Darko y Gul 2005).

Para llevar a cabo las pruebas unitarias se utilizó el marco de trabajo Groovy JUnit, el cual constituye un entorno para ejecutar pruebas internas en el lenguaje de programación Groovy. Groovy Junit crea automáticamente las clases en las cuales se implementarán los métodos para realizar las pruebas

unitarias ya sean a las clases de dominio o controladoras. Estas clases se encuentran dentro del directorio *test*, en la carpeta *unit*.

Se definieron ocho pruebas unitarias para las funciones principales que contiene la clase controladora MetadataExtraction en tres iteraciones, ejemplo: *save()* y *delete()*. Cada uno de los métodos permite comprobar si al ejecutarse las funciones, estas hacen correctamente la tarea para la cual fueron creadas y si a partir de una entrada determinada se obtiene la salida deseada.

En una primera iteración de pruebas resultaron insatisfactorias seis pruebas unitarias, los errores encontrados fueron: mal redireccionamiento desde una acción a otra en el controlador, valores nulos que no eran validados por los métodos y otros errores a causa de una mala implementación de las pruebas unitarias, no se verificaba la salida del método que se estaba probando. Resueltos los errores presentados en la primera iteración, en la segunda fueron satisfactorias cinco pruebas unitarias y el resto incorrectas. El principal error obtenido en estas tres pruebas está relacionado a la incorrecta implementación de las pruebas unitarias, el objeto creado de la clase de dominio MetadataExtraction se tomaba como valor nulo por lo que no se podía validar correctamente. Corregidos los errores obtenidos anteriormente se ejecuta la tercera y última iteración, donde fueron ejecutadas correctamente todas las pruebas unitarias.

En la **Figura 7** se muestran parte de los resultados obtenidos una vez concluida la primera iteración de pruebas unitarias. En la **Figura 8** se presenta la salida obtenida a partir del marco de trabajo JUnit al corregir todos los errores encontrados. La ejecución de las pruebas unitarias permitió validar si el código programado está listo, funcional y cumple con los requisitos funcionales satisfaciendo así las necesidades del cliente y de los usuarios finales.

```
|Configuring classpath
.
|Environment set to test
.............................................
|Running without daemon...
| Compiling 1 source files
| Compiling 1 source files.
| Running 5 unit tests...
| Running 5 unit tests... 1 of 5
| Running 5 unit tests... 2 of 5
| Running 5 unit tests... 3 of 5
| Failure:  Test the save action correctly persists an instance(sdl.extraction.MetadataExtractionControllerSpec)
|  Condition not satisfied:
model.metadataExtraction != null
|     |                    |
|     null                 false
[extractionInstance:sdl.extraction.MetadataExtraction : (unsaved)]
    at sdl.extraction.MetadataExtractionControllerSpec.Test the save action correctly persists an instance(MetadataExtractionControllerSpec.groovy:45)
| Running 5 unit tests... 4 of 5
| Failure:  Test that the show action returns the correct model(sdl.extraction.MetadataExtractionControllerSpec)
|  Condition not satisfied:
response.redirectedUrl == '/extraction/extraction'
|        |             |
|        |             false
|        |             8 differences (73% similarity)
|        |             /(m)e(tadataE)xtraction/extraction
|        |             /(-)e(-------)xtraction/extraction
|        /metadataExtraction/extraction
org.codehaus.groovy.grails.plugins.testing.GrailsMockHttpServletResponse@27a99bd8
    at sdl.extraction.MetadataExtractionControllerSpec.Test that the show action returns the correct model(MetadataExtractionControllerSpec.groovy:66)
| Running 5 unit tests... 5 of 5
| Failure:  Test that the edit action returns the correct model(sdl.extraction.MetadataExtractionControllerSpec)
|  org.codehaus.groovy.grails.web.servlet.mvc.exceptions.CannotRedirectException: Cannot issue a redirect(..) here. A previous call to redirect(..) has already redirected the
 response.
    at sdl.extraction.MetadataExtractionController.$tt__edit(MetadataExtractionController.groovy:57)
    at sdl.extraction.MetadataExtractionControllerSpec.Test that the edit action returns the correct model(MetadataExtractionControllerSpec.groovy:89)
```

Figura 7: Resultados de la iteración uno de las pruebas unitarias

```
/usr/lib/jvm/java-7-openjdk-amd64/bin/java -Dgrails.home=/home/leduan/Apps/grails-2.5.2 -Dbase.dir=/home/leduan/Development/sdl
  .jar=/usr/lib/jvm/java-7-openjdk-amd64/lib/tools.jar -Dgroovy.starter.conf=/home/leduan/Apps/grails-2.5.2/conf/groovy-starter
  -XX:PermSize=256m -Dfile.encoding=UTF-8 -classpath /home/leduan/Apps/grails-2.5.2/lib/org.codehaus.groovy/groovy-all/jars/gro
  .2/dist/grails-bootstrap-2.5.3.jar org.codehaus.groovy.grails.cli.support.GrailsStarter --main org.codehaus.groovy.grails.cli
  .2/conf/groovy-starter.conf "test-app -plain-output"|Loading Grails 2.5.3
|Configuring classpath
.
|Environment set to test
.........................................
|Running without daemon...
| Running 5 unit tests...
| Running 5 unit tests... 1 of 5
| Running 5 unit tests... 2 of 5
| Running 5 unit tests... 3 of 5
| Running 5 unit tests... 4 of 5
| Running 5 unit tests... 5 of 5
| Running 5 unit tests... 6 of 6
| Running 5 unit tests... 7 of 7
| Running 5 unit tests... 8 of 8
| Completed 8 unit tests, 0 failed in 0m 26s
```

Figura 8: Resultados obtenidos al concluir la tercera iteración de pruebas unitarias

3.2.1.2. Prueba de integración

Una vez realizadas las pruebas unitarias se procede a ejecutar las pruebas de integración. Estas pruebas se encargan de validar el correcto funcionamiento en conjunto de todos los módulos que componen un sistema de software. En este trabajo se lleva a cabo esta prueba con el objetivo de comprobar si la herramienta GROBID ha sido integrada correctamente al componente para la extracción de metadatos bibliográficos. Para realizar este tipo de prueba se utiliza el marco de trabajo Groovy JUnit.

La ejecución de la prueba de integración dio como resultados que la herramienta GROBID se integra correctamente al componente. Esto significa que una vez que los artículos científicos son almacenados en el directorio de documentos estos pasan a ser procesados directamente por la herramienta GROBID sin ningún inconveniente. En la figura siguiente se muestra la salida obtenida una vez terminada la ejecución de la prueba en el marco de trabajo JUnit:

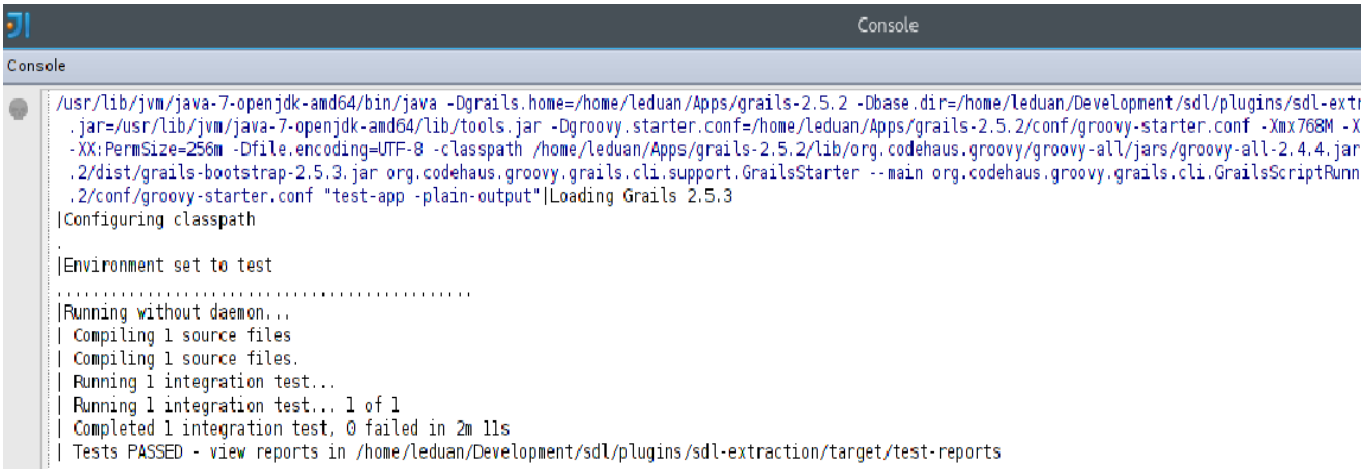


Figura 9: Resultados prueba de integración

3.2.1.3. Pruebas de caja negra

En la prueba de caja negra, se utilizan casos de prueba para demostrar que las funciones del software son correctas, que la entrada se acepta de forma adecuada y que se produce una salida esperada. Las pruebas se llevan a cabo sobre la interfaz visual del sistema. No enfocan su atención en cómo se generan las respuestas del sistema. Básicamente el enfoque de este tipo de prueba se basa en el análisis de los datos de entrada y en los de salida. Los casos de prueba de caja negra pretenden demostrar que:

- Las funciones del software son operativas
- La entrada se acepta de forma correcta
- Se produce una salida correcta
- La integridad de la información externa se mantiene

Existen varios tipos de pruebas de caja negra, entre las que se encuentran, (Blanco Bueno 2016):

- Prueba de Partición Equivalente: divide el dominio de entrada de un programa en clases de datos, a partir de las cuales deriva los casos de prueba
- Prueba de Análisis de Valores Límites: se basa en la evidencia experimental de que los errores suelen aparecer con mayor probabilidad en los extremos de los campos de entrada

De los tipos de pruebas antes mencionadas se decide utilizar la prueba de **Partición Equivalente**, ya que, a partir de un conjunto de datos de entrada, que conforman los Casos de Prueba (CP), se definen estados válidos y no válidos del sistema. Para la validación de la solución utilizando las pruebas de Caja negra se diseñaron cinco CP, correspondientes a los requisitos funcionales *Introducir datos y*

documentos y *Catalogar documentos*, ya que estos cuentan con interfaces visuales. Se detalla, a continuación, el CP correspondiente al requisito *Introducir datos y documentos*.

Caso de prueba: Introducir datos del Evento y documentos

Este CP se encarga de validar la interfaz visual para el requisito funcional Introducir datos y documentos. El usuario tiene la posibilidad en el CP de introducir una colección de documentos o un documento perteneciente a un Evento, así como el número de edición del Evento. En el CP se validan también las entradas incorrectas, o sea, si no se selecciona el Tipo de colección o no se guardan documentos en la aplicación para ser procesados.

El CP está compuesto por seis escenarios, donde se validan cada uno de los campos que componen la interfaz de usuario. El escenario número seis le permite al usuario no realizar ninguna operación, es decir, seleccionar el botón Cancelar que lo redirecciona a la interfaz principal de la aplicación. La tabla con la descripción del CP: Introducir datos del Evento y documentos aparece en el Anexo 2.

Condiciones de ejecución

- El usuario debe seleccionar el componente **Extracción**.
- El usuario debe seleccionar la opción **Extracción de metadatos**.

Tabla 10: Descripción de las variables utilizadas en el caso de pruebas: Introducir documentos

No	Nombre de campo	Clasificación	Valor Nulo	Descripción
1	Tipo de colección	Campo de búsqueda	No	Selección
2	Nombre de la colección	Campo de búsqueda	No	Selección
3	Edición de la colección	Campo de búsqueda	No	Selección
4	Insertar artículos en formato PDF	Campo de entrada de documentos	No	Entrada de archivos

En las **Figura 10** se muestra la ejecución del *CP: Introducir datos del Evento y documentos* en su escenario 1.3, una vez resueltas las no conformidades encontradas en la primera iteración de pruebas. Cada uno de estos escenarios representan flujos de trabajo incorrectos, ya que el usuario no selecciona el Tipo de colección, no guarda los documentos en la aplicación o los documentos seleccionados por él no están en formato PDF. En las figuras se pueden apreciar también los errores que muestra la aplicación si el usuario no cumple con el flujo de funcionamiento.

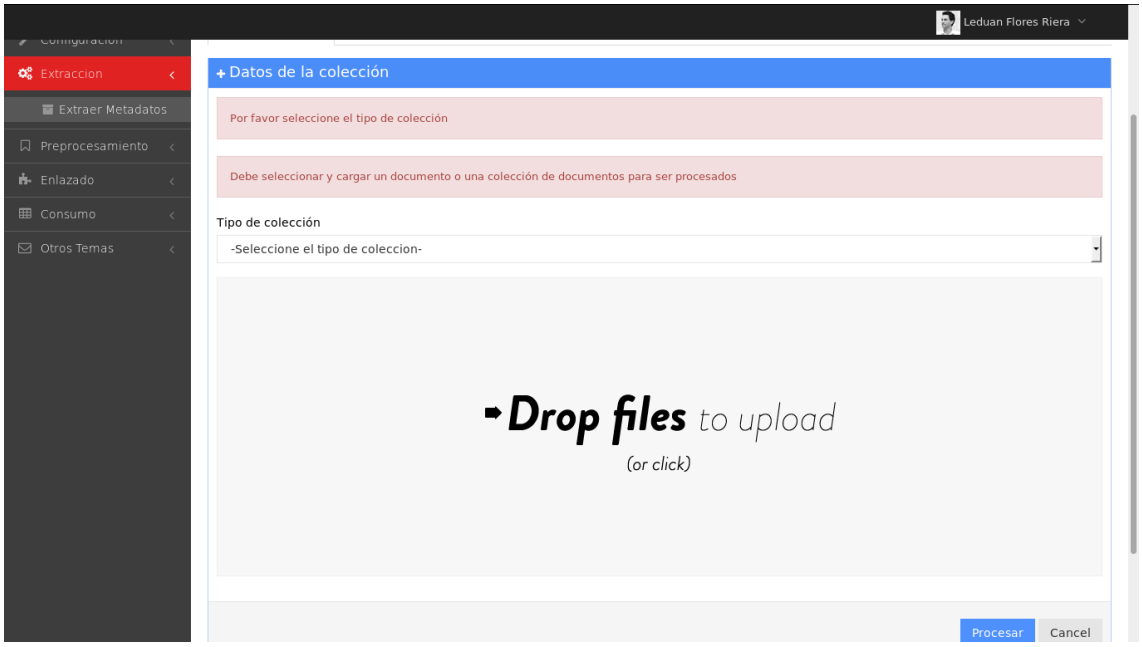


Figura 10: CP Introducir datos del Evento y documentos. Escenario 1.3

En las **Tabla 11** y **Tabla 12** se muestran los casos de prueba *Mostrar metadatos actualizados* y *Seleccionar documento a Catalogar*. Estos CP solo describen en sus escenarios el flujo que debe seguir la aplicación para mostrar la información en la interfaz de usuario y qué debe suceder cuando el usuario presiona el botón Cancelar.

Tabla 11: CP Mostrar metadatos actualizados

Escenario	**Descripción**	**Metadatos**	**Respuesta del sistema**	**Flujo central**
EC 1.1 Se muestra una tabla con los metadatos actualizados y se presiona el botón Listo.	Se muestra una interfaz con los metadatos actualizados por los usuarios.	V Metadatos	El sistema muestra una interfaz con los metadatos del documento ya actualizados.	- Se muestra la interfaz con los metadatos ya actualizados - El usuario presiona el botón **Listo**. **-** Los metadatos son guardados. - El sistema redirecciona al usuario a la interfaz principal de la opción **Extraer metadatos**.
		V		

EC 1.2 Se muestra una tabla con los metadatos actualizados y se presiona el botón Cancelar	Se muestra la interfaz inicial de la opción **Extraer metadatos**.	N/A	El sistema redirecciona al usuario a la interfaz principal de la opción **Extraer metadatos**.	´- Se presiona el botón **Cancelar**. - El sistema redirecciona al usuario a la interfaz principal de la opción **Extraer metadatos**.

Tabla 12: Seleccionar documentos a catalogar

Escenario	**Descripción**	**Nombre del documento**	**Respuesta del sistema**	**Flujo central**
EC 1.1 Existen documentos para catalogar se selecciona el documento que desea catalogar dando click sobre el nombre del documento.	Se muestra una interfaz con el listado de documentos a catalogar. Al seleccionar el documento que se desea catalogar el sistema le muestra una interfaz con el documento y sus metadatos	V Documento	El sistema muestra una interfaz con el documento seleccionado y sus metadatos.	- Se selecciona la pestaña **Documentos Procesados**. - Se muestra la interfaz correspondiente con el listado de documentos a catalogar. - El usuario selecciona el documento que desea catalogar dando click sobre el nombre el documento PDF. - El sistema muestra una interfaz con el documento seleccionado y sus metadatos.

EC 1.2 No hay documentos para catalogar.	Se muestra una interfaz sin el listado de documentos.	V N/A	El sistema muestra la interfaz correspondiente a la pestaña Documentos Procesados sin el listado de documentos a procesar	- Se selecciona la pestaña **Documentos Procesados**. - Se muestra la interfaz correspondiente sin el listado de documentos a catalogar.
EC 1.3 Existen documentos para catalogar.	Se muestra una interfaz sin el listado de documentos.	I N/A		- Se selecciona la pestaña **Documentos Procesados**. - Se muestra la interfaz correspondiente sin el listado de documentos a catalogar.

Resultados

Con los diseños de casos de pruebas creados se procede a la ejecución de las pruebas. En una primera iteración de los cinco casos de pruebas propuestos tres resultaron no satisfactorios, representando el 60% del total.

Las no conformidades encontradas fueron referidas a que la aplicación permitía dejar campos sin seleccionar o vacíos. Se encontraron no conformidades sobre errores ortográficos, la internacionalización de los textos y mensajes que se presentan en la aplicación, estos no se mostraban correctamente. En total se encontraron 20 no conformidades de ellas 15 no conformidades de interfaz y cinco de funciones incorrectas. Culminada la primera iteración de pruebas se procedió a corregir las no conformidades encontradas. En una segunda iteración todos los CP resultaron satisfactorios.

3.2.2. Pruebas de aceptación con el cliente

Para realizar las pruebas de aceptación se utilizó la técnica de validación con el cliente. El cliente es quien valida si la aplicación está lista para ser utilizada por los usuarios finales. Para la validación del software se utilizaron los casos de pruebas definidos para las pruebas funcionales, además de un conjunto de datos de prueba que son las entradas a cada uno de los CP.

3.3. Caso de estudio

Con el objetivo de validar la solución al problema de investigación se diseña un caso de estudio. Se utiliza para ello una colección de 200 artículos en formato PDF los cuales están almacenados *a priori* en un directorio local y posteriormente estos son incorporados al servidor de la aplicación para ser procesados. La colección de artículos científicos proviene de las memorias del evento Informática 2013. Para el caso de estudio se cuenta con un equipo de cómputo con las siguientes prestaciones:

- Tipo de CPU: Intel Dual Core 2.10 GHz
- Memoria del sistema: 3 Gb de RAM

3.4. Diseño experimental

Se utiliza en la investigación un pre-experimento para validar la propuesta de solución. Para el pre-experimento se precisa del resultado de una observación inicial que será comparada en otro momento con los valores obtenidos luego de la aplicación de un estímulo. Se definen cuatro tareas a realizar, enumeradas seguidamente:

- Procesar 10 documentos en formato PDF.
- Procesar 50 documentos en formato PDF.
- Procesar 100 documentos en formato PDF.
- Procesar 200 documentos en formato PDF.

En la tabla siguiente se muestra el diseño experimenta propuesto:

Tabla 13: Diseño experimental

Muestra	Tareas	Observación simple	Estímulo	Observación estímulo
10 AC-PDF	TiPx	OSi	CEMB	OEi

- **TiPx**: Tarea i que realiza el especialista en bibliotecología, Procesar x cantidad de artículos científicos en formato PDF.
- **OSi**: Observación simple, tiempo que demora el especialista en realizar la tarea **Ti**.
- **OEi**: Observación del estímulo, tiempo que demora el CEMB en realizar la **Ti.**
- **AC-PDF** (Artículo Científicos en formato PDF)
- **CEMB** (Componente para la Extracción de Metadatos Bibliográficos)

Se proponen los siguientes escenarios para la evaluación, en cada uno de ellos se medirá el tiempo que demora la extracción de los metadatos bibliográficos:

1. Realizar la extracción de los metadatos bibliográficos de artículos científicos en formato PDF de manera manual, sin la utilización de la propuesta de solución.
2. Extraer los metadatos bibliográficos de artículos científicos en formato PDF utilizando la propuesta de solución como estímulo.

3.5. Análisis de los resultados

Una vez realizada la medición del tiempo que demoran los especialistas en extraer los metadatos bibliográficos a diez artículos científicos se obtiene un tiempo medio de *2:08.80* minutos por artículo. El proceso de extracción utilizando el estímulo, el Componente para la Extracción de Metadatos Bibliográficos (CEMB), se obtiene un tiempo promedio de *1:53.60* minutos por documento. En la **Tabla 14** se muestra el diseño experimental propuesto y se aplican los resultados obtenidos para determinar cuánto demorarían los especialistas y el CEMB en el procesado de varias cantidades de artículos científicos, en este caso desde 10 hasta 200 artículos científicos. La población utilizada para realizar el pre-experimento es de 200 artículos científicos.

Tabla 14: Diseño experimental propuesto

Fuentes de datos	Tareas	Observación simple	Estímulo	Observación estímulo
200 PDF	***T1:*** Procesar 10 PDF	***OS1:*** 00:20:80	**CEMB**	***OE1:*** 00:15:60
	T2: Procesar 50 PDF	***OS2:*** 01:47:30		***OE2:*** 01:16:08
	T3: Procesar 100 PDF	***OS3:*** 03:34:66		***OE3:*** 02:33:06
	T3: Procesar 200 PDF	***OS4:*** 07:09:33		***OE4:*** 05:07:20

GROBID presenta un alto desempeño en el procesamiento de los artículos científicos. Según los creadores de la herramienta para una colección de 4000 PDF GROBID realiza el proceso de extracción

de metadatos del encabezado de los documentos en 10 minutos, o sea, 3 PDF por segundo y 18 segundos procesando 3000 referencias bibliográficas.[29]

La aplicación del CEMB reduce el tiempo en aproximadamente 55 segundos y dos centésimas siendo una solución factible para ser introducida dentro de un ambiente real donde uno de sus procesos sea la extracción de metadatos bibliográficos. En el análisis de este resultado se debe tener en cuenta la disponibilidad de recursos de hardware donde es puesto en funcionamiento el componente, ya que este proceso requiere de un alto procesamiento.

En (Carr y Harnad 2005) se plantea que el tiempo medio que demora una persona en llevar a cabo el proceso de extracción de metadatos es de 5 minutos y 37 segundos por artículo científico. El CEMB reduce el tiempo de extracción de metadatos bibliográficos de artículos científicos por una persona planteado en el artículo en aproximadamente en 3 minutos y 44 segundos. El CEMB es una solución viable para llevar a cabo el proceso de extracción de metadatos bibliográficos.

3.6. Conclusiones parciales

Con la aplicación de las pruebas unitarias se detectaron errores en el código implementado que a simple vista no se habían detectado. Las pruebas de caja negra se realizaron con el cliente interactuando con la aplicación a partir de un flujo definido en los casos de pruebas, dando como resultado las no conformidades que surgieron durante el proceso. En el diseño experimental se demostró que el Componente para la extracción de metadatos bibliográficos reduce el tiempo que demoran los especialistas en bibliotecología en extraer los metadatos bibliográficos a partir de artículos científicos en formato PDF. Una vez validado el Componente para la extracción de metadatos bibliográficos se dice que este es una solución factible para ser aplicada en una biblioteca para realizar uno de sus procesos claves, la extracción de metadatos, así como para ser integrado al proyecto Extracción, publicación y consumo de metadatos bibliográficos como datos enlazados en la web, como uno de sus componentes.

[29] Ver, https://github.com/kermitt2/grobid

CONCLUSIONES GENERALES

1. La revisión de la literatura evidenció que las bibliotecas digitales son un elemento importante para lograr la búsqueda y el acceso a la información científica. Esto se logra a través de los metadatos los cuales identifican a un documento científico dentro de la biblioteca digital y posibilitan tener un mayor y rápido acceso al conocimiento.
2. El estudio de las principales aproximaciones permitió conocer cuáles son las herramientas dedicadas a la extracción automática de metadatos, así como los métodos que utilizan para el procesamiento de los documentos. Se analizaron también los lenguajes y estándares para la representación de los metadatos, siendo el lenguaje XML el más utilizado. A partir de este estudio se determinó que GROBID tiene un mejor desempeño que otras herramientas en la extracción de metadatos bibliográficos.
3. El componente se encarga de extraer los metadatos bibliográficos a cada uno de los documentos guardados en la aplicación y luego esos metadatos pasan por un proceso de revisión para comprobar su calidad, llevado a cabo por el usuario. Está basado en una arquitectura de tuberías y filtros.
4. La herramienta GROBID fue integrada a la solución propuesta por sus resultados en el análisis realizado en el estado del arte de la investigación. GROBID está implementada en el lenguaje Java. Este elemento es importante para la integración con la solución, al estar desarrollada utilizando Groovy, como lenguaje de la Java Virtual Machine puede ser integrado fácilmente con aplicaciones Java.
5. La propuesta de solución disminuye el tiempo que consume un especialista en realizar el proceso de extracción de metadatos bibliográficos en una biblioteca. Se dice que es una solución factible para ser aplicada en una biblioteca digital, así como para ser integrada al proyecto Extracción, publicación y consumo de metadatos bibliográficos como datos enlazados en la web, como uno de sus componentes.

RECOMENDACIONES

- Los modelos de GROBID no están entrenados para procesar artículos científicos en idioma español, por lo que se recomienda realizar un entrenamiento de la herramienta con colecciones de documentos en idioma español.
- Actualmente el componente implementado solo está diseñado para procesar artículos científicos publicados en revistas y eventos. Se recomienda extender las funcionalidades del componente para extraer metadatos de otros documentos científicos tales como Libros y Tesis.

REFERENCIAS BIBLIOGRÁFICAS

BERNERS-LEE, T., HENDLER, J. y LASSILA, O., 2001. The Semantic Web. *Scientific American*, pp. 29-37.

BLANCO BUENO, C., 2016. *Ingeniería del Software I. Construcción y Pruebas de Software* [en línea]. 2016. S.l.: s.n. Disponible en: http://ocw.unican.es/ensenanzas-tecnicas/ingenieria-del-software-ii/materiales.

CARR, L. y HARNAD, S., 2005. Keystroke Economy: A Study of the Time and Effort Involved in Self-Archiving. , pp. 1-7.

CHAVE, M.A., 2007. La ingeniería de requerimientos y su importancia en el desarrollo de proyectos de software. [en línea], vol. VI, no. 10. ISSN 1409-4746. Disponible en: http://revistas.ucr.ac.cr/index.php/intersedes/article/view/790.

COUNCILL, I.G., LEE GILES, C. y KAN, M.-Y., 2015. ParsCit: An open-source CRF reference string parsing package. ,

FENN, J., 2006. Managing Citations and Your Bibliography with BibTEX. , no. 4, pp. 1-19.

FLYNN, P.K., 2014. *Document Classification in Support of Automated Metadata Extraction from Heterogeneous Collections*. Doctorado. S.l.: Old Dominion University,.

GRANITZER, M., MAYA, H. y ROBERT, K., 2015. A Comparison of Metadata Extraction Techniques for Crowdsourced Bibliographic Metadata Management. ,

GRANITZER, M., MAYA, H., ROBERT, K., JACK, K. y KERN, R., 2012. A Comparison of Layout based Bibliographic Metadata Extraction Techniques. *2nd International Conference on Web Intelligence, Mining and Semantics*. S.l.: s.n., ISBN 978-1-4503-0915-8. DOI 10.1145/2254129.2254154.

GREENBERG, J., 2009. Understanding Metadata and Metadata Schemes. *Cataloging & Classification Quarterly*, vol. 40, pp. 17-36. DOI 10.1300/J104v40n03_02.

GREENBERG, J., SPURGIN, K. y CRYSTAL, A., 2006. Functionalities for automatic metadata generation applications: a survey of metadata experts' opinions. *Int. J. Metadata, Semantics and Ontologies*, vol. 1, no. 1, pp. 3-20.

GREENBERG, J., SWAUGER, S. y FEINSTEIN, E.M., 2013. Metadata Capital in a Data Repository. *Dublin Core and Metadata Applications 201*. S.l.: s.n., pp. 140-150.

HIDALGO DELGADO, Y. y RODRÍGUEZ PUENTE, R., 2013. La web semántica: una breve revisión. , vol. 7, no. 1, pp. 76-85. ISSN 2227-1899.

HÍPOLA, P., VARGAS-QUESADA, B. y A. SENSO, J., 2000. Bibliotecas digitales: situación actual y problemas. *El Profesional de la información*, vol. 9, no. 4, pp. 4-13.

HUI HAN, LEE GILES, C., MANAVOGLU, E. y ZHA, H., 2003. Automatic Document Metadata Extraction using Support Vector Machines. *Proceedings of the 2003 Joint Conference on Digital Libraries (JCDL'03)*, DOI 0-7695-1939-3/03.

HUSBY, O., 1997. Metadata: Elag'97. ,

JOHNSTON, P., 2005. Metadata Sharing and XML. [en línea], Disponible en: http://www.ukoln.ac.uk/nof/support/help/papers/metaxml/index.html.

KHOO, M., PARK, ung-R. y XIA, L., 2009. THE USER-CENTERED DESIGN OF A NON-SPECIALIST METADATA TOOL AND INTERFACE FOR THE INTERNET PUBLI C LIBRARY. . S.l.: s.n., pp. 1-2.

KOUSHIK, S., DARKO, M. y GUL, A., 2005. CUTE: a concolic unit testing engine for C. [en línea]. S.l.: s.n., pp. 63-272. ISBN 1-59593-014-0. DOI 10.1145/1081706.1081750. Disponible en: http://dl.acm.org/citation.cfm?id=1081750.

LAFFERTY, J., MCCALLUM, A. y PEREIRA, F.C.N., 2001. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. *Eighteenth International Conference on Machine Learning* [en línea]. S.l.: s.n., pp. 282-289. ISBN 1-55860-778-1. Disponible en: http://portal.acm.org/citation.cfm?id=655813.

L. BORGMAN, C., 1999. What are digital libraries? Competing visions. , vol. 35, pp. 227-243.

LIDDY, E.D., SUTTON, S., PAIK, W., ALLEN, E., HARWELL, S., MONSOUR, M., TURNER, A. y LIDDY, J., 2001. Breaking the Metadata Generation Bottleneck: Preliminary Findings. . S.l.: s.n., pp. 464. DOI 1-58113-345-6/01/0006.

LIPINSKI, M., KEVIN, Y., JOERAN, B. y BELA, G., 2013. Evaluation of Header Metadata Extraction Approaches and Tools for Scientific PDF Documents. , pp. 385-386.

LÓPEZ, P., 2015. GROBID: Combining Automatic Bibliographic Data Recognition and Term Extraction For Scholarship Publications. ,

LÓPEZ, P. y ROMARY, L., 2015. HUMB: Automatic Key Term Extraction from Scientific Articles in GROBID. ,

LO RUSSO, G., SPOLVERI, F., CIANCIO, F. y MORI, A., 2013. Mendeley: An Easy Way to Manage, Share, and Synchronize Papers and Citations. *Plastic and Reconstructive Surgery*, pp. 946-947. DOI 10.1097/PRS.0b013e31828bd400.

LÖSCH, U., BLOEHDORN, S. y RETTINGER, A., 2012. Graph kernels for RDF data. En: 00053, *The Semantic Web: Research and Applications*. S.l.: Springer, pp. 134–148.

MILLER, P., 1996. Metadata for the mases. *Ariadne*, no. 5.

MITCHELL, T.M., 1997. *Machine Learning*. S.l.: McGraw-Hill Science/Engineering/Math. ISBN 0-07-042807-7.

MITCHELL, T.M., 2006. The Discipline of Machine Learning. ,

PINILLA, A., GUTIÉRREZ, M. y BALLEJOS, L., 2014. EXTRACCIÓN AUTOMÁTICA DE METADATOS A PARTIR DE OBJETOS DE APRENDIZAJE EN UN REPOSITORIO INSTITUCIONAL: ESTADO DEL ARTE. , pp. 67-82. ISSN 2362-5139.

PRESSMAN, R.S., 2006. *Ingeniería del software: Un enfoque práctico*. 5. S.l.: s.n.

PYTEL, P., UHALDE, C., RAMÓN, H., CASTELLO, H., TOMASELLO, M., POLLO-CATTANEO, M., BRITOS, P. y GARCÍA MARTÍNEZ, R., 2011. INGENIERÍA DE REQUISITOS BASADA EN TÉCNICAS DE IN GENIERÍA DEL CONOCIMIENTO. *XIII Workshop de Investigadores en Ciencias de la Computación*. S.l.: s.n., pp. 426-429. ISBN 978-950-673-892-1.

RODRÍGUEZ SÁNCHEZ, T., 2015. *Metodología de desarrollo para la Actividad productiva UCI*. 6 marzo 2015. S.l.: s.n.

SÁNCHEZ DÍAZ, M. y VEGA VALDÉS, J.C., 2002. Bibliotecas electrónicas, digitales y virtuales: tres entidades por definir. [en línea], vol. 10, no. 2. ISSN 1024-9435. Disponible en: http://scielo.sld.cu/.

SCHAPIRE, R., 2008. *Theoretical Machine Learning* [en línea]. 4 febrero 2008. S.l.: s.n. Disponible en: http://www.cs.princeton.edu/courses/archive/spr08/cos511/.

SENSO, J.A. y DE LA ROSA PIÑERO, A., 2003. El concepto de metadato. Algo más que descripción de recursos electrónicos. , vol. 32.

SICILIA, M.-A., 2014a. *HANDBOOK OF METADATA, SEMANTICS AND ONTOLOGIES*. S.l.: World Scientific Publishing. ISBN 978-981-283-629-8.

SICILIA, M.-A., 2014b. *Handbook of metadata, semantics and ontologies*. S.l.: World Scientific.

SUTTON, C. y MCCALLUM, 2012. An Introduction to Conditional Random Fields. , vol. 4, pp. 267-373. DOI 10.1561/2200000013.

TESTA, P. y CERIOTTO, P., 2011. Descripción de objetos digitales: metadatos. En: 00002, *Encuentro Nacional de Catalogadores (2do: 2009: Buenos Aires, Argentina)*, vol. 1, pp. 105–112.

TRAMULLAS SAZ, J., 2012. Propuestas de concepto y definición de la biblioteca digital. ,

VÁSQUEZ PAULUS, C., 2015. *METADATOS: Introducción e historia*. S.l.: s.n.

WALLACH, H.M., 2014. Conditional Random Fields: An Introduction. , pp. 1-9.

ZHANG, F. y ZHAO, Z., 2013. A Metadata Extraction Approach from Papers Based on Meta-learning⋆. En: 00000,

ANEXOS

Anexo 1 CP: Introducir datos de la Revista y documentos en formato PDF

Escenario	Descripción	Tipo de colección	Nombre de la colección	Volumen de la colección	Número de la colección	Insertar artículos en formato PDF	Respuesta del sistema	Flujo central
EC 1.1 Seleccionar todos los datos y los documentos en formato PDF. Presionar el botón Procesar	Los documentos en formato PDF son guardados en el directorio de documentos.	V Revista	V Nombre	V Volumen	V Número	V Documentos en PDF	El sistema guarda los documentos insertados en el directorio de documentos.	- Se selecciona la opción **Extraer metadatos** en la parte izquierda de la interfaz. - Se muestra la interfaz correspondiente. - El usuario selecciona el Tipo de colección, Nombre de la colección, el Volumen y el Número o la Edición de la colección. - El usuario selecciona el documento o una colección de documentos en formato PDF para ser guardados. - El sistema comprueba si el formato de los documentos es PDF. - El usuario presiona el botón **Procesar**. - El sistema valida si todos los datos han sido seleccionados y si los documentos han sido guardados. - El sistema procede a extraer los metadatos bibliográficos de los documentos introducidos.

EC 1.2 Seleccionar todos los datos y no introducir los documentos en formato PDF.	Los documentos en formato PDF no son guardados en el directorio de documentos.	V Revista	V Nombre	V Volumen	V Número	I Documentos en PDF	El sistema no guarda los documentos insertados en el directorio de documentos. Se muestra el mensaje de error "Debe seleccionar un documento o una colección de documentos para ser procesados".	- Se selecciona la opción **Extraer metadatos** en la parte izquierda de la interfaz. - Se muestra la interfaz correspondiente. - El usuario selecciona el Tipo de colección, Nombre de la colección, el Volumen y el Número o la Edición de la colección. - El usuario no selecciona el documento o una colección de documentos en formato PDF para ser guardados. - El usuario presiona el botón **Procesar**. - El sistema valida si todos los datos han sido seleccionados y si los documentos han sido guardados. - El sistema muestra el mensaje de error "Debe seleccionar un documento o una colección de documentos para ser procesados".
EC 1.3 Dejar de seleccionar algunos de los datos y no introducir los documentos en formato PDF.	Los documentos en formato PDF no son guardados en el directorio de documentos.	I N/A	V No se muestra	V No se muestra	V No se muestra	I Documentos en PDF	El sistema no guarda los documentos insertados en el directorio de documentos. Se muestra un mensaje de error por cada dato no seleccionado y el mensaje "Debe seleccionar un documento o una colección de documentos para ser procesados".	- Se selecciona la opción **Extraer metadatos** en la parte izquierda de la interfaz. - Se muestra la interfaz correspondiente. - El usuario no selecciona el Tipo de colección o el Nombre de la colección o el Volumen o el Número o la Edición de la colección. - El usuario no selecciona el documento o una colección de documentos en formato PDF para ser guardados. -El usuario presiona el botón **Procesar**. - El sistema valida si todos los datos han sido seleccionados y si los documentos han sido guardados. - El sistema muestra un mensaje de error por cada dato no seleccionado y el mensaje "Debe seleccionar un documento o una colección de documentos para ser procesados"

EC 1.4 Dejar de seleccionar algunos de los datos e introducir los documentos en formato PDF.	Los documentos en formato PDF son guardados en el directorio de documentos.	V Revista	I N/A	V No se muestra	V No se muestra	V Documentos en PDF	El sistema guarda los documentos insertados en el directorio de documentos. Se muestra un mensaje de error por cada dato no seleccionado.	- Se selecciona la opción **Extraer metadatos** en la parte izquierda de la interfaz. - Se muestra la interfaz correspondiente. - El usuario no selecciona el Tipo de colección o el Nombre de la colección o el Volumen o el Número o la Edición de la colección. - El usuario selecciona el documento o una colección de documentos en formato PDF para ser guardados. - El sistema comprueba si el formato de los documentos es PDF. - El usuario presiona el botón **Procesar**. - El sistema valida si todos los datos han sido seleccionados y si los documentos han sido guardados. - El sistema muestra un mensaje de error por cada dato no seleccionado.
EC 1.5 Seleccionar los datos e introducir los documentos en un formato diferente a PDF u otro tipo de archivo.	Los documentos no son guardados en el directorio de documentos.	V Revista	V Nombre	V Volumen	V Número	I Documentos en otro formato	El sistema no guarda los documentos insertados en el directorio de documentos. Se señaliza con una cruz que los archivos introducidos no son los correctos.	- Se selecciona la opción **Extraer metadatos** en la parte izquierda de la interfaz. - Se muestra la interfaz correspondiente. - El usuario selecciona el Tipo de colección o el Nombre de la colección o el Volumen o el Número o la Edición de la colección. - El usuario selecciona el documento o una colección de documentos en un formato diferente a PDF u otro tipo de archivo para ser guardados. - El sistema comprueba si el formato de los documentos es PDF. - El sistema indica con una cruz que el documento no está en el formato correcto.

EC 1.6 Presionar el botón Cancelar	Se muestra la interfaz principal de la aplicación.	V N/A	V N/A	V N/A	V N/A	V N/A	El sistema redirecciona al usuario a la interfaz principal de la aplicación	- Se presiona el botón Cancelar. - El sistema redirecciona al usuario a la interfaz principal de la aplicación

Descripción de las variables

No	Nombre de campo	Clasificación	Valor Nulo	Descripción
1	Tipo de colección	Campo de búsqueda	No	Selección
2	Nombre de la colección	Campo de búsqueda	No	Selección
3	Volumen de la colección	Campo de búsqueda	No	Selección
4	Número de la colección	Campo de búsqueda	No	Selección
5	Insertar artículos en formato PDF	Campo de entrada de documentos	No	Entrada de archivos

Anexo 2 CP: Introducir datos del Evento y documentos en formato PDF

Escenario	Descripción	Tipo de colección	Nombre de la colección	Edición de la colección	Insertar artículos en formato PDF	Respuesta del sistema	Flujo central
EC 1.1 Seleccionar todos los datos y los documentos en formato PDF. Presionar el botón Procesar	Los documentos en formato PDF son guardados en el directorio de documentos.	V Evento	V Nombre	V Edición	V Documentos en PDF	El sistema guarda los documentos insertados en el directorio de documentos.	- Se selecciona la opción **Extraer metadatos** en la parte izquierda de la interfaz. - Se muestra la interfaz correspondiente. - El usuario selecciona el Tipo de colección, Nombre de la colección, el Volumen y el Número o la Edición de la colección. - El usuario selecciona el documento o una colección de documentos en formato PDF para ser guardados. - El sistema comprueba si el formato de

							los documentos es PDF. - Presiona el botón **Procesar.** **-** El sistema valida si todos los datos han sido seleccionados y si los documentos han sido guardados. - El sistema procede a extraer los metadatos bibliográficos de los documentos introducidos.
EC 1.2 Seleccionar todos los datos y no introducir los documentos en formato PDF.	Los documentos en formato PDF no son guardados en el directorio de documentos.	V Evento	V Nombre	V Edición	I Documentos en PDF	El sistema no guarda los documentos insertados en el directorio de documentos. Se muestra el mensaje de error "Debe seleccionar un documento o una colección de documentos para ser procesados".	- Se selecciona la opción **Extraer metadatos** en la parte izquierda de la interfaz. - Se muestra la interfaz correspondiente. - El usuario selecciona el Tipo de colección, Nombre de la colección, el Volumen y el Número o la Edición de la colección. - El usuario no selecciona el documento o una colección de documentos en formato PDF para ser guardados. - Presiona el botón **Procesar**. - El sistema valida si todos los datos han sido seleccionados y si los documentos han sido guardados. - El sistema muestra el mensaje de error "Debe seleccionar un documento o una colección de documentos para ser procesados".
EC 1.3 Dejar de seleccionar algunos de los datos y no introducir los documentos en formato PDF.	Los documentos en formato PDF no son guardados en el directorio de documentos.	I N/A	V No se muestra	V No se muestra	I Documentos en PDF	El sistema no guarda los documentos insertados en el directorio de documentos. Se muestra un mensaje de error por cada dato no seleccionado y el mensaje "Debe	- Se selecciona la opción **Extraer metadatos** en la parte izquierda de la interfaz. - Se muestra la interfaz correspondiente. - El usuario no selecciona el Tipo de colección o el Nombre de la colección o el Volumen o el Número o la Edición de la colección. - El usuario no selecciona el documento o una colección de documentos en formato PDF para ser guardados. - Presiona el botón **Procesar**. - El sistema valida si todos los datos han sido seleccionados y si los documentos

						seleccionar un documento o una colección de documentos para ser procesados".	han sido guardados. - El sistema muestra un mensaje de error por cada dato no seleccionado y el mensaje "Debe seleccionar un documento o una colección de documentos para ser procesados"
EC 1.4 Dejar de seleccionar algunos de los datos e introducir los documentos en formato PDF.	Los documentos en formato PDF son guardados en el directorio de documentos.	V Evento	I N/A	V No se muestra	V Documentos en PDF	El sistema guarda los documentos insertados en el directorio de documentos. Se muestra un mensaje de error por cada dato no seleccionado.	- Se selecciona la opción **Extraer metadatos** en la parte izquierda de la interfaz. - Se muestra la interfaz correspondiente. - El usuario no selecciona el Tipo de colección o el Nombre de la colección o el Volumen o el Número o la Edición de la colección. - El usuario selecciona el documento o una colección de documentos en formato PDF para ser guardados. - El sistema comprueba si el formato de los documentos es PDF. - Presiona el botón **Procesar**. - El sistema valida si todos los datos han sido seleccionados y si los documentos han sido guardados. - El sistema muestra un mensaje de error por cada dato no seleccionado y el mensaje "Debe seleccionar un documento o una colección de documentos para ser procesados"
EC 1.5 Seleccionar los datos e introducir los documentos en un formato diferente a PDF u otro tipo de archivo.	Los documentos no son guardados en el directorio de documentos.	V Revista	V Nombre	V Edición	I Documentos en otro formato	El sistema no guarda los documentos insertados en el directorio de documentos. Se señaliza con una cruz que los	- Se selecciona la opción **Extraer metadatos** en la parte izquierda de la interfaz. - Se muestra la interfaz correspondiente. - El usuario selecciona el Tipo de colección o el Nombre de la colección o el Volumen o el Número o la Edición de la colección. - El usuario selecciona el documento o una colección de documentos en un formato

						archivos introducidos no son los correctos.	diferente a PDF u otro tipo de archivo para ser guardados. - El sistema comprueba si el formato de los documentos es PDF. - El sistema indica con una cruz que el documento no está en el formato correcto.
EC 1.6 Presionar el botón Cancelar	Se muestra la interfaz principal de la aplicación.	V N/A	V N/A	V N/A	V N/A	El sistema redirecciona al usuario a la interfaz principal de la aplicación	- Se presiona el botón **Cancelar**. - El sistema redirecciona al usuario a la interfaz principal de la aplicación.

Descripción de las variables

No	Nombre de campo	Clasificación	Valor Nulo	Descripción
1	Tipo de colección	Campo de búsqueda	No	Selección
2	Nombre de la colección	Campo de búsqueda	No	Selección
3	Edición de la colección	Campo de búsqueda	No	Selección
4	Insertar artículos en formato PDF	Campo de entrada de documentos	No	Entrada de archivos

Anexo 3 CP: Catalogar metadatos del documento seleccionado

Leyenda:

T: Título; **N**: Nombre; **No.**: Número; **F**: Fecha; el resto lo puede visualizar en la tabla de la descripción de las variables.

Escenario	Descripción	TA	A	AF	PC	ID	F	PI	PF	FP	FR	Respuesta del Sistema	Flujo Central
		V	V	V	V	V	V	V	V	V	V		

EC 1.1 Catalogar los metadatos bibliográficos del documento seleccionado, llenando cada uno de los campos requeridos correctamente. Presionar el botón Actualizar.	Se muestra una interfaz con el documento y sus metadatos bibliográficos. El usuario modifica los metadatos que considere incorrectos y presiona el botón **Actualizar**.	T	N	AF	PC	ID	PDF	No.	No.	F	F	El sistema actualiza los metadatos del documento seleccionado en la base de datos y muestra una interfaz con los metadatos del documento ya actualizados.	- Se selecciona la pestaña **Documentos procesados**. - Se muestra la interfaz correspondiente. - El usuario selecciona el documento que desea catalogar dando click sobre el nombre. - Se muestra una interfaz con el documento seleccionado y una tabla con los metadatos extraídos. - El usuario modifica los metadatos que tengan errores y agrega los que falten. - Presiona el botón **Actualizar**. **-** El sistema valida si no hay campo vacíos y si los números de páginas son correctos. - Si todos los campos han sido llenados se actualizan los metadatos en la base de datos. - El sistema muestra una interfaz con los metadatos ya actualizados.
EC 1.2 Catalogar los metadatos bibliográficos del documento seleccionado, dejando campos vacíos. Presionar el botón Actualizar.	Se muestra una interfaz con el documento y sus metadatos bibliográficos. El usuario no modifica todos los metadatos y presiona el botón **Actualizar**.	V T	V N	V AF	I PC	V ID	V PDF	V No.	V No.	I F	I F	El sistema no actualiza los metadatos del documento seleccionado en la base de datos y muestra un error indicando que se deben llenar todos los campos.	´- Se selecciona la pestaña **Documentos procesados**. - Se muestra la interfaz correspondiente. - El usuario selecciona el documento que desea catalogar dando click sobre el nombre. - Se muestra una interfaz con el documento seleccionado y una tabla con los metadatos extraídos.

													- El usuario deja campos vacíos. - Presiona el botón **Actualizar**. **-** El sistema valida si no hay campo vacíos y si los números de páginas son correctos.. - Si hay campos vacíos el sistema se mantiene en la misma interfaz y muestra un mensaje de error indicando que se deben llenar todos los campo.
EC 1.3 Catalogar los metadatos bibliográficos del documento seleccionado, introduciendo el número de página de inicio mayor que el número de página final. Presionar el botón Actualizar.	Se muestra una interfaz con el documento y sus metadatos bibliográficos. El usuario introduce valores incorrectos en los números de páginas y presiona el botón **Actualizar**.	V T	V N	V AF	V PC	V ID	V PDF	I No.	I No.	V F	V F	El sistema no actualiza los metadatos del documento seleccionado en la base de datos y muestra el mensaje de error "El número de la página de inicio debe ser menor que el número de la página final".	- Se selecciona la pestaña **Documentos procesados**. - Se muestra la interfaz correspondiente. - El usuario selecciona el documento que desea catalogar dando click sobre el nombre. - Se muestra una interfaz con el documento seleccionado y una tabla con los metadatos extraídos. - El usuario introduce valores incorrectos los números de páginas. - Presiona el botón **Actualizar**. **-** El sistema valida si no hay campo vacíos y si los números de páginas son correctos. - Si el número de página inicial es mayor que el número de página final el sistema muestra el mensaje de error "El número de la página de inicio debe ser

													menor que el número de la página final" y se mantiene en la misma interfaz.
EC 1.4 Presionar el botón Cancelar	Se muestra la interfaz inicial de la opción **Extraer metadatos**.	V N/A	V N/A	V N/A	V N/A	V N/A	V N/A	V N/A	V N/A	V N/A	V N/A	El sistema redirecciona al usuario a la interfaz principal de la opción **Extraer metadatos**.	- Se presiona el botón **Cancelar**. - El sistema redirecciona al usuario a la interfaz principal de la opción **Extraer metadatos**.

Descripción de las variables

No	Nombre de campo	Clasificación	Valor Nulo	Descripción
1	Título del artículo (TA)	Campo de texto	No	Texto
2	Autores (A)	Campo de texto	No	Texto
3	Afiliación (AF)	Campo de texto	No	Texto
4	Palabras clave (PC)	Campo de texto	No	Texto
5	Idioma (ID)	Campo de búsqueda	No	Selección
6	Formato (F)	Campo de búsqueda	No	Selección
7	Página de inicio (PI)	Campo de texto	No	Texto
8	Página de fin (PF)	Campo de texto	No	Texto
9	Fecha de publicación (FP)	Campo de fecha	No	Selección de fecha
10	Fecha de recepción (FR)	Campo de fecha	No	Selección de fecha