Revista Cubana de Ciencias Informáticas
Vol. 11, No. 4, Octubre-Diciembre, 2017
ISSN: 2227-1899 | RNPS: 2301
http://rcci.uci.cu
Pág. 85- 98

Tipo de artículo: Artículo original
Temática: Inteligencia Artificial
Recibido: 17/03/2017 | Aceptado: 15/10/2017

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

Component for automatic metadata extraction from textual corpus in PDF

Leduan Flores Riera[1*], Alejandro Mariño Molerio[1], Luis Mojena Román[1], Yusniel Hidalgo Delgado[1]

[1] Universidad de las Ciencias Informáticas. Carretera a San Antonio de los Baños, km 2 ½, Lisa, Ciudad de La Habana, Cuba. lflores@uci.cu, ajmarino@uci.cu, lamojena@estudiantes.uci.cu, yhdelgado@uci.cu

[*] Autor para correspondencia: lflores@uci.cu

Resumen

Las bibliotecas digitales se encargan de la gestión documental de los recursos digitales que almacenan, realizando tres procesos fundamentales: la selección, tratamiento y explotación de los recursos. La extracción de los metadatos es una de las tareas del tratamiento de los documentos digitales, facilita la búsqueda, acceso y recuperación de la información. La extracción de metadatos es un proceso que requiere tiempo para su ejecución y en caso de ejecutarse manualmente puede existir el riesgo de introducir errores humanos. Estos problemas se pueden aliviar con el uso de herramientas automatizadas que apoyen esta actividad. En este artículo se describe 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 una arquitectura de tuberías y filtros, donde la salida de un proceso constituye la entrada al próximo. Para validar si el componente de extracción de metadatos reduce el tiempo de extracción se realiza un diseño experimental a partir de un caso de estudio. Además de validar el componente a través del diseño experimental se le aplican un conjunto de pruebas de calidad. Estas pruebas van encaminadas a comprobar si el funcionamiento del componente es el adecuado, si las funciones implementadas se ejecutan correctamente, si los resultados obtenidos son los deseados y si el usuario final tiene un nivel alto de aceptación con el componente de extracción de metadatos.

Palabras clave: Artículos científicos, Documentos PDF, Extracción de metadatos, Metadatos, Web Semántica.

Abstract

Digital libraries are responsible for management of stored digitals resources and perform three fundamental processes: the selection, treatment and exploitation of resources. One of the functions of treatment is the metadata extraction process; in order to facilitate its use, that is, allow the search, access and retrieval of information.

Revista Cubana de Ciencias Informáticas
Vol. 11, No. 4, Octubre-Diciembre, 2017
ISSN: 2227-1899 | RNPS: 2301
http://rcci.uci.cu
Pág. 85- 98

Metadata extraction is a process that requiring time for its execution and if executed manually could there is the risk of introducing human errors. These problems can be reduced by the use of automated tools to support this process. In this article, we describe a web component for automatic extraction of bibliographic metadata from PDF files. The component is based on three fundamental processes that follow a data flow that represents a tubes and filters architecture, where the output of one process constitutes the input to the next. To validate if the metadata extraction component reduces the extraction time, an experimental design is made using a case study. Furthermore, a set of quality tests is applied. These tests are aimed at verifying if the functioning of the component is correct, if the implemented functions are executed correctly, if the obtained results are the desired ones and if the user has a high level of acceptance with the component of extraction of metadata.

Keywords: *Scientific articles, Metadata extraction, Metadata, PDF Documents, Semantic Web.*

Introducción

La Web Semántica surge con el objetivo de resolver las limitaciones: integración, formato y recuperación de la web actual y como una extensión de esta. 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 et al. 2001; Wenger 2014). 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 de computación que surgen para apoyar el trabajo realizado en las bibliotecas físicas, llevando a cabo la gestión documental de los recursos u objetos digitales que almacenan. Este proceso consta de la selección, tratamiento y explotación de los documentos, donde en el tratamiento se realizan un conjunto de tareas como la catalogación y extracción de los metadatos de los documentos, libros y otros recursos, con el objetivo de que los usuarios puedan acceder más rápido a la información que buscan y tener almacenados los datos que identifican a cada objeto contenido en la biblioteca ya sea digital o físico.

El proceso de extracción de metadatos se encarga de obtener los atributos o etiquetas que identifican a cada documento (Senso y Piñero 2003; Nogueira 2013). Estos metadatos servirán para la búsqueda, recuperación, autenticación y evaluación de un recurso dentro de la biblioteca digital. Realizar este proceso manualmente requiere de expertos en bibliotecología y puede demorar teniendo en consideración la cantidad de documentos a los cuales se les extraerán los metadatos (Flynn 2014). Como vía de solución se han desarrollado aplicaciones que pueden ser utilizadas desde la web o como una aplicación de escritorio. 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.

La extracción de metadatos es un proceso que requiere tiempo y se lleva a cabo para identificar y extraer los metadatos como son el título y los autores, para luego ser guardados en una base de datos en línea. 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 2013). 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 2009; Cerejo 2013).

En este artículo se describe un componente de software para la extracción de metadatos bibliográficos a partir de documentos en formato PDF. Con la utilización de este componente se podría reducir el tiempo empleado por los especialistas en bibliotecología para la extracción de metadatos bibliográficos. El componente ha sido desarrollado para ser desplegado en un servidor web, por lo que les brinda la ventaja a los usuarios de tener acceso desde cualquier computadora, siempre que exista una conexión a internet. Además, al estar desarrollado en forma de componente favorece su reutilización en otros proyectos donde se utilicen metadatos bibliográficos. Para la implementación del componente se emplearon herramientas de código abierto o libre lo que reduce los costos durante la etapa de desarrollo del componente.

Materiales y Métodos

En este acápite se realiza un análisis sobre los tipos de metadatos existentes según la bibliografía consultada, donde se especifican sus aplicaciones y se dan ejemplos de estos tipos de metadatos. Para la extracción de metadatos bibliográficos se han desarrollado herramientas, las cuales pueden ser utilizadas desde la web, como aplicaciones de escritorio o pueden ser integradas a otros proyectos. De las herramientas existentes en este apartado se lleva a cabo un estudio sobre tres herramientas con el objetivo describir sus principales características y determinar sus ventajas.

Tipos de metadatos existentes

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 2013). A continuación, se explican tres tipos de metadatos existente en la literatura consultada:

Los **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 Piñero 2003; Testa 2013). Mientras que, los **metadatos administrativos** 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 Piñero 2003; Testa 2013). Por último, los **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 2012; Testa 2013). 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.

Herramientas para la extracción de metadatos bibliográficos

Existen varias herramientas dedicadas a la extracción de metadatos bibliográficos de documentos científicos y técnicos en formato PDF, de las cuales se seleccionaron aquellas que utilizan técnicas de aprendizaje automático tales como: Grobid, Mendeley y ParsCit. A continuación, se caracterizan cada una de ellas.

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 (Lopez y Romary 2010a; Hasan y Ng 2014). Es software libre, desarrollado utilizando el lenguaje de programación Java. Puede ser utilizada como una aplicación web o integrada a otros sistemas. Puede extraer metadatos bibliográficos tales como: autores, el título, el resumen, palabras claves y otros. Para lograr el reconocimiento de la estructura del documento y la extracción de los metadatos la herramienta realiza la conversión de los documentos científicos en formato PDF a documentos en formato TEI (Text Encoding Initiative) (Lopez y Romary 2010b; Hasan y Ng 2014). Los metadatos extraídos pueden ser representados utilizando BibTex, lenguaje para la descripción de bibliografía. La herramienta se enfoca en las secciones: encabezado (título, resumen), introducción, la sección de títulos, las conclusiones y las referencias bibliográficas, ya que en estas secciones los autores introducen los conceptos principales y los lectores suelen prestar más atención a estas partes del documento. Su uso puede ser extendido a las bibliotecas digitales como un módulo para el análisis y procesado de documentos de texto, esto permite la obtención de información para generar y sugerir citas bibliográficas a los usuarios (Lopez 2009; Tkaczyk et al. 2015).

Revista Cubana de Ciencias Informáticas
Vol. 11, No. 4, Octubre-Diciembre, 2017
ISSN: 2227-1899 | RNPS: 2301
http://rcci.uci.cu
Pág. 85- 98

Mendeley

Software libre, que combina un sitio web y una aplicación para PC y dispositivos Apple (IPhone y IPad) para el almacenamiento y manejo de documentos PDF. Permite tener los documentos almacenados en la Nube y también compartirlos con otros como una red social. La aplicación organiza automáticamente los artículos por categorías (autor, título, revista, fecha y demás) en una base de datos para luego realizar filtrados por categorías. Proporciona el manejo de referencias bibliográficas, la selección o creación de estilos de citas textuales y la creación automática de bibliografía (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 (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, Giles y Kan 2008; Guy et al. 2014; Ramakrishnan et al. 2012). En el preprocesado, ParsCit utiliza métodos heurísticos para convertir el documento en formato PDF a texto plano, empleando UTF-8 (Councill, Giles y Kan 2008; Guy et al. 2014; Ramakrishnan et al. 2012). 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 et al. 2012). La herramienta puede ser utilizada tanto como un servicio web o como una aplicación independiente.

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.

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 1.

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

Revista Cubana de Ciencias Informáticas
Vol. 11, No. 4, Octubre-Diciembre, 2017
ISSN: 2227-1899 | RNPS: 2301
http://rcci.uci.cu
Pág. 85- 98

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.

Tabla 1. 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

En el artículo (Granitzer et al. 2012) 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.

Resultados y discusión

A continuación, se describe el 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 comparación hecha en el acápite anterior entre las herramientas descritas arrojó como resultado que Grobid es la que mejor desempeño tiene en el proceso de extracción de metadatos bibliográficos. Por esta razón se decide utilizar Grobid para ser integrada a la propuesta de solución.

El componente es descrito a través del diagrama de procesos mostrado en la figura 1.

En el diagrama de procesos se visualizan seis subprocesos que en su conjunto conforman el proceso de extracción de metadatos. De estos subprocesos se describen a continuación tres de ellos, ya que son los de mayor relevancia para la implementación del componente:

Revista Cubana de Ciencias Informáticas
Vol. 11, No. 4, Octubre-Diciembre, 2017
ISSN: 2227-1899 | RNPS: 2301
http://rcci.uci.cu
Pág. 85- 98

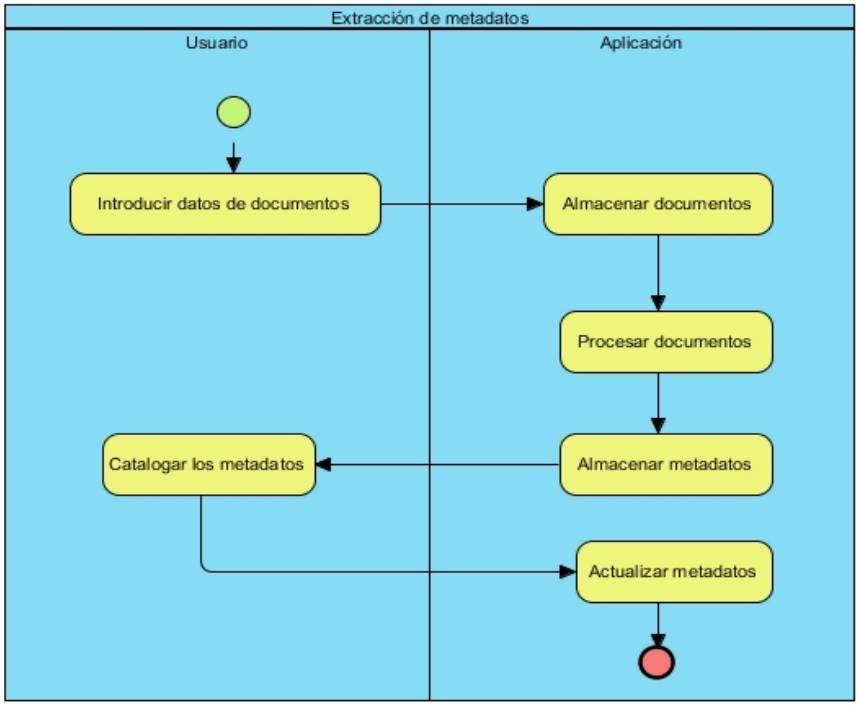

Figura 1: Diagrama de procesos del componente

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, la cual 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 mismo. Los metadatos pueden ser editados si están incorrectos y se actualizan directamente en la base de datos.

Arquitectura del componente

El componente 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 y Maxim 2015).

En la **Figura 2** se muestra el diseño arquitectónico del componente. 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.

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 incluye 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.

Revista Cubana de Ciencias Informáticas
Vol. 11, No. 4, Octubre-Diciembre, 2017
ISSN: 2227-1899 | RNPS: 2301
http://rcci.uci.cu
Pág. 85- 98

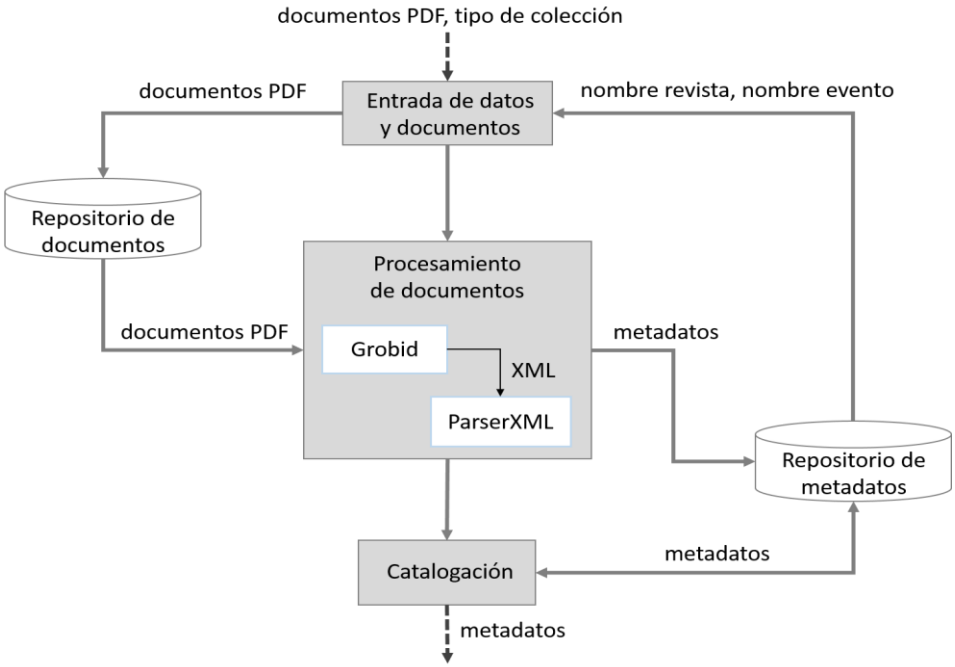

Figura 2: Arquitectura del componente

En la **Figura 3** se puede observar una captura del componente en funcionamiento. La imagen corresponde al filtro catalogación de metadatos:

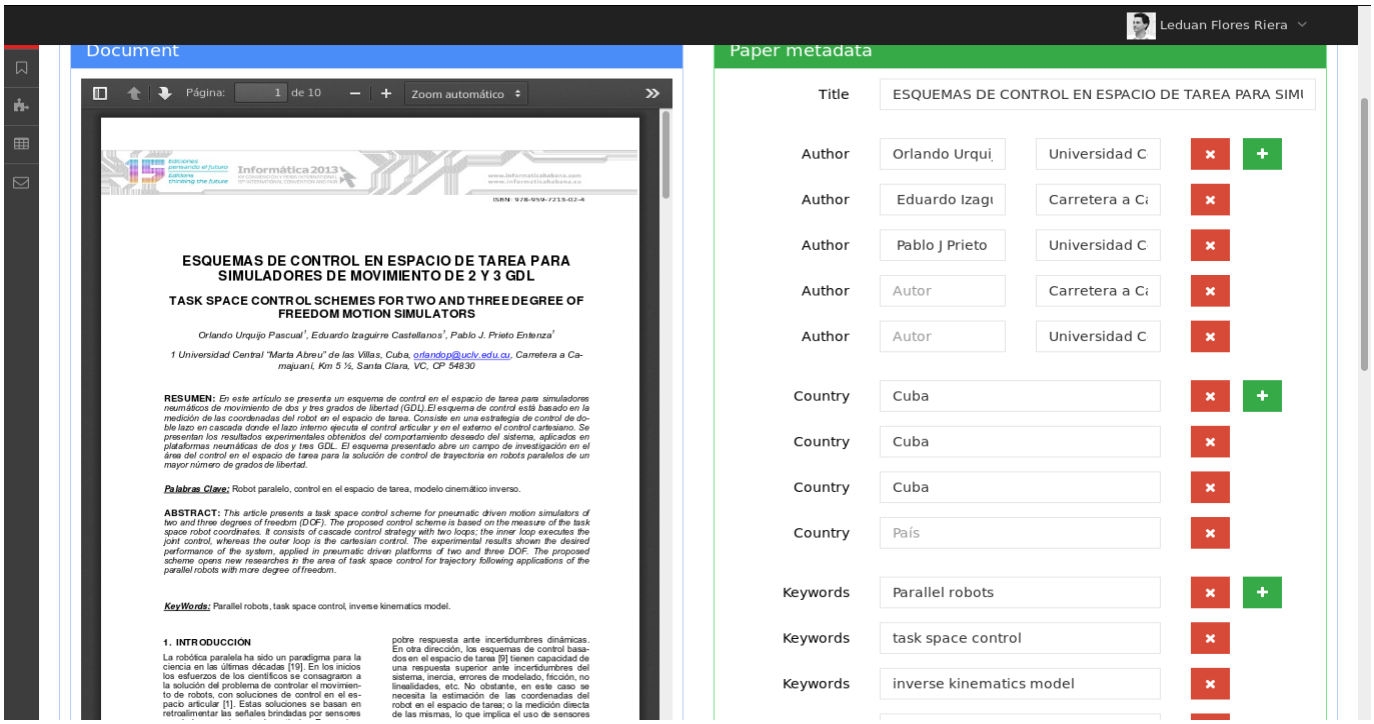

Figura 3: Captura de pantalla del componente para la extracción de metadatos bibliográficos

Revista Cubana de Ciencias Informáticas
Vol. 11, No. 4, Octubre-Diciembre, 2017
ISSN: 2227-1899 | RNPS: 2301
http://rcci.uci.cu
Pág. 85- 98

Validación de resultados

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

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 2: 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 definieron los siguientes escenarios para la evaluación, en cada uno de ellos se midió 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.

Revista Cubana de Ciencias Informáticas
Vol. 11, No. 4, Octubre-Diciembre, 2017
ISSN: 2227-1899 | RNPS: 2301
http://rcci.uci.cu
Pág. 85- 98

2. Extraer los metadatos bibliográficos de artículos científicos en formato PDF utilizando la propuesta de solución como estímulo.

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 3** 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 3: 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
	T4: 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 (Lopez 2017).

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 desplegado el componente, ya que este proceso requiere de un alto procesamiento.

En (Carr y Harnad 2009; Cerejo 2013) 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.

Como parte del proceso de validación del componente se realizaron pruebas de calidad para comprobar su ejecución, si los resultados obtenidos eran los deseados y si el cliente estaba de acuerdo con la herramienta. Las pruebas aplicadas fueron: unitarias, de integración, de caja negra y de aceptación con el cliente. En cada una de las pruebas

implementadas se realizaron entre una y tres iteraciones hasta obtener el resultado correcto. La aplicación de las pruebas permitió la detección de errores en el código implementado que a simple vista no se habían detectado. Además, las pruebas realizadas con el cliente interactuando con la aplicación a partir de un flujo definido en los casos de pruebas, dieron como resultados las no conformidades que surgieron durante el proceso y permitieron conocer el nivel de aceptación que tenía el cliente con el componente desarrollado.

Conclusiones

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. Al reducirse el tiempo de extracción de metadatos se da cumplimiento al objetivo trazado inicialmente en el artículo. Con el desarrollo de este experimento se pudo demostrar la aplicabilidad y factibilidad del componente para ser adoptado en una biblioteca para realizar procesos de extracción de metadatos. Actualmente el componente implementado solo está diseñado para procesar artículos científicos publicados en revistas y eventos. Se está trabajando en extender las funcionalidades del componente para extraer metadatos de otros documentos científicos tales como Libros y Tesis.

Referencias

BERNERS-LEE, T., HENDLER, J., LASSILA, O. y OTHERS, 2001. The semantic web.

CARR, L. y HARNAD, S., 2009. Keystroke Economy: A Study of the Time and Effort Involved in Self-Archiving. 2005. *Web (Accessed:)*,

CEREJO, C., 2013. How to make your paper more accessible through self-archiving. *Editage Insights (04-11-2013)* [en línea], Disponible en: goo.gl/Lm95X3.

COUNCILL, I.G., GILES, C.L. y KAN, M.-Y., 2008. ParsCit: an Open-source CRF Reference String Parsing Package. . S.l.: s.n.,

FLYNN, P.K., 2014. *Document Classification in Support of Automated Metadata Extraction from Heterogeneous Collections*. S.l.: OLD DOMINION UNIVERSITY.

GRANITZER, M., HRISTAKEVA, M., JACK, K. y KNIGHT, R., 2012. A comparison of metadata extraction techniques for crowdsourced bibliographic metadata management. *Proceedings of the 27th Annual ACM Symposium on Applied Computing*. S.l.: ACM, pp. 962–964.

GUY, T., GLASZIOU, P., CHOONG, M.K., DUNN, A., GALGANI, F. y COIERA, E., 2014. Systematic review automation technologies. *BioMed Central*, vol. 3, no. 1, pp. 74. DOI 10.1186/2046-4053-3-74.

HASAN, K.S. y NG, V., 2014. Automatic Keyphrase Extraction: A Survey of the State of the Art. *ACL (1)*. S.l.: s.n., pp. 1262–1273.

HIDALGO DELGADO, Y. y RODRÍGUEZ PUENTE, R., 2013. La web semántica: una breve revisión. *Revista Cubana de Ciencias Informáticas*, vol. 7, no. 1, pp. 76–85.

LIPINSKI, M., YAO, K., BREITINGER, C., BEEL, J. y GIPP, B., 2013. Evaluation of header metadata extraction approaches and tools for scientific PDF documents. *Proceedings of the 13th ACM/IEEE-CS joint conference on Digital libraries*. S.l.: ACM, pp. 385–386.

LOPEZ, P., 2009. GROBID: Combining automatic bibliographic data recognition and term extraction for scholarship publications. *International Conference on Theory and Practice of Digital Libraries*. S.l.: Springer, pp. 473–474.

LOPEZ, P., 2017. *grobid: A machine learning software for extracting information from scholarly documents* [en línea]. Java. S.l.: s.n. [Consulta: 16 junio 2017]. Disponible en: https://github.com/kermitt2/grobid.

LOPEZ, P. y ROMARY, L., 2010a. HUMB: Automatic key term extraction from scientific articles in GROBID. *Proceedings of the 5th international workshop on semantic evaluation*. S.l.: Association for Computational Linguistics, pp. 248–251.

LOPEZ, P. y ROMARY, L., 2010b. HUMB: Automatic key term extraction from scientific articles in GROBID. *Proceedings of the 5th international workshop on semantic evaluation*. S.l.: Association for Computational Linguistics, pp. 248–251.

NOGUEIRA, D.M., 2013. *Herramientas de apoyo a la Gestión por el Conocimiento para docentes e investigadores de las Ciencias Empresariales en Cuba.* S.l.: s.n.

PRESSMAN, R.S. y MAXIM, B.R., 2015. *Software Engineering: A Practitioner's Approach*. S.l.: s.n. ISBN 0-07-802212-6.

RAMAKRISHNAN, C., PATNIA, A., HOVY, E. y BURNS, G.A., 2012. Layout-aware text extraction from full-text PDF of scientific articles. *Source Code for Biology and Medicine*, vol. 7, no. 1, pp. 7. ISSN 1751-0473. DOI 10.1186/1751-0473-7-7.

RUSSO, G.L., SPOLVERI, F., CIANCIO, F. y MORI, A., 2013. Mendeley: An easy way to manage, share, and synchronize papers and citations. *Plastic and reconstructive surgery*, vol. 131, no. 6, pp. 946e–947e.

SENSO, J.A. y PIÑERO, A. de la R., 2003. El concepto de metadato. Algo más que descripción de recursos electrónicos. *Ciência da Informação*, vol. 32, no. 2, pp. 95–106.

SICILIA, M.-A., 2013. *Handbook of metadata, semantics and ontologies*. S.l.: World Scientific.

TESTA, P., 2013. *Esquemas de metadatos para los repositorios institucionales de las universidades nacionales argentinas*. S.l.: s.n.

TESTA, P. y CERIOTTO, P., 2012. Descripción de objetos digitales: metadatos. *Sistema Integrado de Documentación, Universidad Nacional del Cuyo*,

TKACZYK, D., SZOSTEK, P., FEDORYSZAK, M., DENDEK, P.J. y BOLIKOWSKI, \Lukasz, 2015. CERMINE: automatic extraction of structured metadata from scientific literature. *International Journal on Document Analysis and Recognition (IJDAR)*, vol. 18, no. 4, pp. 317–335. ISSN 1433-2825. DOI 10.1007/s10032-015-0249-8.

WENGER, E., 2014. *Artificial intelligence and tutoring systems: computational and cognitive approaches to the communication of knowledge*. S.l.: Morgan Kaufmann.