Revista Publicando, 5. 14 (3). 2018, 571-582. ISSN 1390-9304

Herramienta para el procesamiento en lote de documentos digitales para el sistema REPXOS 3.0.

José Javier Hernández Benítez[1], Aray Villar Machado[2], Deylert Pérez Rivera[3], Rafael Alejandro Linares Torres[4]

1 Universidad de las Ciencias Informáticas, jjbenitez@uci.cu

2 Universidad de las Ciencias Informáticas, aray@uci.cu

3 Universidad de las Ciencias Informáticas, drivera@uci.cu

4 Universidad de las Ciencias Informáticas, ralinres@uci.cu

RESUMEN

REPXOS es un sistema para la implantación de repositorios digitales en instituciones científicas y académicas, desarrollado en el centro de Informatización de la Gestión Documental (CIGED), perteneciente a la Universidad de las Ciencias Informáticas (UCI). Para almacenar la información en REPXOS es necesario llenar de forma manual documento a documento un formulario con los metadatos que lo describe. Este proceso es fácil de realizar cuando es poca información, pero cuando es en el orden de los cientos o miles de documentos, trae consigo pérdida de tiempo y empleo de recursos humanos, por lo que en ocasiones no se incorporan los documentos más antiguos, provocando que se pierda parte del patrimonio académico y científico de las instituciones. Por esta situación se decidió desarrollar la aplicación web DEPXOS con el objetivo de extraer metadatos automáticamente de documentos digitales para su posterior depósito por lotes en el sistema REPXOS. Se utilizará para la investigación los métodos teóricos analítico-sintético y la modelación y como método empírico la observación. El resultado de esta investigación es una herramienta informática capaz de extraer los metadatos de varios documentos digitales, realizar su validación basados en el parámetro de completitud y realizar su depósito en el sistema REPXOS posteriormente.

Palabras claves: documentos digitales, herramienta informática, información, metadatos.

Recibido 03/03/2018
Aprobado 09/04/2018

Revista Publicando, 5. 14 (3). 2018, 571-582. ISSN 1390-9304

Tool for batch processing of digital documents for the REPXOS 3.0 system.

ABSTRACT

REPXOS 3.0 is a system developed in the University of Computer Science (UCI), specifically in the Center for the Computerization of Document Management (CIGED), for setting up digital repositories in scientific and academic institutions, to store the information in REPXOS 3.0 it is necessary to introduce manually a metadata in a form with its description, document by document. This process is easy to perform when it is little information, but when it is in the order of the hundreds or thousands of documents, it brings about a loss of time and use of human resources, that is the reason why sometimes the older documents are not incorporated and also part of the academic and scientific heritage of the institutions is lost. For this situation it was decided to develop the web application DEPXOS with the objective of extracting metadata automatically from digital documents for later deposit in the system REPXOS 3.0. The theoretical analytic and synthetic method and the modeling method will be used for the investigation and observation as empirical method. The result of this research is a computer tool capable of extracting metadata from several digital documents, performing their validation based on the parameter of completeness and making their deposit in the REPXOS 3.0 system at a later time.

Keywords: computer tool, digital documents, information, metadata.

1. INTRODUCCIÓN

Recibido 03/03/2018
Aprobado 09/04/2018

Revista Publicando, 5. 14 (3). 2018, 571-582. ISSN 1390-9304

Las instituciones académicas y científicas generan gran cantidad de información asociada a artículos científicos, revistas, trabajos de diplomas, libros, tesis de maestrías, tesis de doctorados, eventos, conferencias y clases. Toda esta información es de vital importancia almacenarla, con el objetivo de preservar y consultar el patrimonio científico de estas instituciones.

Las Tecnologías de la Información y las Comunicaciones (TIC) proporcionan un conjunto de aplicaciones y sistemas informáticos, que ofrecen la posibilidad de crear bibliotecas y repositorios digitales, con el fin de organizar, preservar, centralizar y difundir el patrimonio científico de las instituciones. Estas aplicaciones informáticas facilitan el almacenamienuncian los métodos de investigación utilizados durante la investigacinto y preservación de los documentos; así como su recuperación a través de herramientas de búsqueda por medio de los metadatos asociados a cada objeto digital.

Un ejemplo es REPXOS 3.0, un sistema de implantación de Repositorios Digitales desarrollado en el Centro de Informatización de la Gestión Documental (CIGED) perteneciente a la Universidad de las Ciencias Informáticas (UCI). Algunas instituciones que hacen uso de este sistema son el Instituto de Geografía Tropical (IGT), el Tribunal Supremo Popular (TSP), el Instituto Superior de Tecnologías y Ciencias Aplicadas (InSTEC) y la Oficina de Asuntos Históricos del Consejo de Estado (OAHCE). El Centro de Información Científico-Técnica de la UCI cuenta también con sus servicios.

Anualmente estas instituciones generan un gran cúmulo de documentos digitales, en el orden de los cientos o miles, que desean incorporar a su Repositorio Digital. Para realizar el depósito de un documento digital en REPXOS 3.0 es necesario ingresar en un formulario, un conjunto de metadatos que describen el mismo. Este proceso se realiza de forma manual por lo que una de las medidas que se toma es designar varios especialistas durante numerosas sesiones de trabajo, para incorporar sus documentos.

Este proceso, aunque es intuitivo, cuando es necesario realizarlo para grandes cantidades de documentos, trae consigo pérdida de tiempo, y empleo de recursos humanos; por lo que en ocasiones los encargados de esta tarea deciden no incluir en el Repositorio Digital los documentos antiguos, perdiéndose así parte del patrimonio científico y académico de las instituciones.

Recibido 03/03/2018
Aprobado 09/04/2018

Revista Publicando, 5. 14 (3). 2018, 571-582. ISSN 1390-9304

Teniendo en cuenta lo antes planteado se propone como objetivo de este trabajo el desarrollo de una herramienta informática que permita la extracción automática de metadatos de documentos en lotes para su posterior depósito en el sistema REPXOS 3.0. Además, el sistema debe permitir validar la calidad de los metadatos de los documentos.

2. METODOS

Métodos Teóricos:

Analítico-Sintético: Utilizado para ejecutar un análisis de la bibliografía correspondiente a la investigación y extraer los elementos significativos para el diseño eficiente de la propuesta de solución.

Modelación: Se utilizó para modelar los diagramas correspondientes a la fase de análisis y diseño, proporcionando una mejor perspectiva de la herramienta a desarrollar.

Métodos Empíricos:

Observación: Se empleó para evaluar los resultados obtenidos contra los resultados esperados con el desarrollo de la herramienta.

3. RESULTADOS

Se realizó un estudio de distintas herramientas para realizar la extracción de metadatos a documentos digitales, basándose en parámetros como la plataforma que soporta, los formatos de documentos digitales que admite y su propósito. A continuación, se muestra un resumen de las distintas herramientas estudiadas.

Apache Tika

Es un conjunto de herramientas para detectar y extraer metadatos y texto estructurado del contenido de varios tipos de documentos usando librerías para el análisis sintáctico. Tika es un proyecto de la Fundación de Software de Apache (ASF) escrito en Java. Proporciona un único conjunto de funciones y una única interfaz Java para acceder a las librerías de análisis sintáctico, consume pocos recursos y posee un rápido procesamiento. Se distribuye bajo la Licencia Apache 2.0. (CHRIS A. MATTMANN, 2012)

Los formatos que puede extraer la herramienta son los siguientes (versión 1.12) (CHRIS A. MATTMANN, 2012)

Recibido 03/03/2018
Aprobado 09/04/2018

Revista Publicando, 5. 14 (3). 2018, 571-582. ISSN 1390-9304

HTML

XML y derivados

Formatos de documentos de Microsoft Office

Formato OpenDocument

Formato de documentos iWorks

Formato de documento portátil (PDF)

Formato de publicaciones electrónicas (EPUB)

Formato de texto enriquecido (RTF)

Formatos de compresión y empaquetado (Tar, AR, CPIO, Zip, 7Zip, Gzip, BZip2, XZ, Pack200 y RAR)

Formato de texto

Formatos de fuentes y sindicación

Formatos de ayuda (CHM)

Formatos de audio

Formatos de imagen

Formatos de videos

Archivos de clases de Java y compilados (JAR)

Código fuente (Java, C, C++, Groovy y otros)

Formatos de correo electrónico

Formatos CAD

Formatos de fuentes tipográficas

Formatos científicos

Programas ejecutables y librerías

Formatos criptográficos

Recibido 03/03/2018
Aprobado 09/04/2018

Revista Publicando, 5. 14 (3). 2018, 571-582. ISSN 1390-9304

Formatos de bases de datos (SQLite y Microsoft Access)

FOCA

Es una herramienta utilizada principalmente para encontrar metadatos e información oculta en los documentos que examina, enfocado en procesos de auditoría de red. Estos documentos pueden estar en páginas web, con FOCA se pueden descargar y analizar. Los documentos son buscados usando tres posibles buscadores: Google, Bing y Exalead, aunque también existe la posibilidad de añadir ficheros locales para extraer sus metadatos. (Eleven Paths, n.d.)

Una vez que se tienen todos los datos extraídos, FOCA une la información buscando qué documentos han sido creados en el mismo equipo y qué servidores y equipos pueden inferirse de dichos datos. FOCA es distribuido bajo la licencia Acuerdo de Licencia con el Usuario Final (EULA). (Eleven Paths, n.d.)

Soporta los formatos (Eleven Paths, n.d.)

Formatos de documentos (Microsoft Office y Open Office)

Formato PDF

Formatos de Adobe InDesing

Formatos SVG

Metadata Extraction Tool

Es una herramienta de código abierto escrita en Java para la extracción de metadatos de distintas fuentes con el objetivo de usarlos para tareas de preservación de recursos digitales. Permite la extracción automática de metadatos enfocados a la preservación de recursos digitales y exporta dichos metadatos en formato XML listos para ser usados en actividades de preservación. Puede ser usada desde una interfaz de usuario en Windows o desde la línea de comandos en sistemas UNIX para procesar ficheros por lotes o de forma individual. Se distribuye bajo los términos de la Licencia Apache 2.0. (Zealand, 2003)

Soporta los formatos (Zealand, 2003)

Formatos de imagen: BMP, GIF, JPEG y TIFF

Recibido 03/03/2018
Aprobado 09/04/2018

Revista Publicando, 5. 14 (3). 2018, 571-582. ISSN 1390-9304

Formatos de documentos (MS Word, Word Perfect, Open Office, MS Works, MS Excel, MS Power Point y PDF)

Formatos de audio y video (WAV y MP3)

Formatos de lenguajes de marcado (HTML y XML)

GROBID

Es una librería de aprendizaje automático de código abierto. Distribuida bajo los términos de la licencia Apache 2.0, para la extracción, análisis sintáctico y la reestructuración de documentos sin estructura, como el formato PDF, en documentos codificados con el estándar TEI con un enfoque particular en publicaciones técnicas y científicas. Posee las siguientes funcionalidades: (GROBID Documentation, 2016)

Extracción de cabecera y análisis sintáctico de artículos en formato PDF. La extracción cubre la información bibliográfica (por ejemplo, título, resumen, autores, afiliaciones, palabras clave, etc.).

Extracción y análisis sintáctico de referencias de artículos en formato PDF. (Funciona con notas al pie de página).

Análisis sintáctico de nombres, particularmente de nombres de autor en las cabeceras y nombres de autor en las referencias (dos modelos distintos).

Análisis sintáctico de afiliaciones y bloques de direcciones.

Análisis sintáctico de fechas.

Extracción completa de texto de artículos en formato PDF.

Fueron seleccionadas para el desarrollo de la aplicación la librería de aprendizaje automático GROBID y el conjunto de herramientas Apache Tika. Con el uso de la funcionalidad para la extracción de cabeceras de GROBID es posible el acceso a información invariable y correcta, algo que se puede ver afectado en los metadatos de los recursos, debido a que estos pueden ser alterados por las aplicaciones donde son creados, editados, etc. Además, GROBID posee un enfoque particular en documentos técnicos y científicos. A pesar de todo lo descrito anteriormente esta librería presenta una desventaja, solo permite el procesamiento de recursos en formato PDF.

Recibido 03/03/2018
Aprobado 09/04/2018

Revista Publicando, 5. 14 (3). 2018, 571-582. ISSN 1390-9304

Debido a que la aplicación se enfrentará a recursos de disímiles formatos se seleccionó también Apache Tika, pues respecto al resto de las herramientas estudiadas es la que mayor cantidad de formatos soporta y posee mecanismos específicos para realizar la extracción de metadatos utilizando el estándar Dublin Core.

Estudio y selección de un método para la validación de metadatos.

La validación de los metadatos extraídos se realizó mediante el uso del marco de trabajo Bruce & Hillman para la evaluación automática de metadatos en repositorios digitales por el hecho de que los siete parámetros utilizados en dicho marco de trabajo son fáciles de entender por humanos y además capturan todas las dimensiones de calidad propuestas en otros marcos de trabajo. Su nivel de compactación también ayuda a realizar la medición de la calidad de los metadatos automáticamente. El marco de trabajo Bruce & Hillman define los siguientes siete parámetros para medir la calidad de los metadatos: (Ochoa & Duval)

Completitud: Una instancia de metadato debe describir el recurso lo más completo posible. Mientras más campos de metadatos contenga el recurso, más diversos serán los usos que se le pueda dar al mismo. (Ochoa & Duval)

Exactitud: La información proporcionada por la instancia de metadato debe ser la más correcta posible. Los errores tipográficos y los errores de hecho afectan esta dimensión de la calidad. La exactitud puede ser medida con valores booleanos en casos como el tamaño del fichero o el formato, sin embargo, campos como el título o la descripción contienen un rango más amplio para evaluar. (Ochoa & Duval)

Procedencia: La fuente de los metadatos puede ser otro factor para determinar su calidad. Conocimiento sobre quién creó la instancia, el nivel de experiencia del indexador, las metodologías que fueron seguidas durante la indexación y las transformaciones que han sufrido los metadatos pueden dar una idea de la calidad de la instancia. (Ochoa & Duval)

Conformidad con lo esperado: El grado con que los metadatos cumplen los requisitos de una determinada comunidad de usuarios para una tarea dada se puede considerar como una dimensión importante de calidad. Si la información almacenada en los metadatos ayuda a la comunidad a encontrar, identificar, seleccionar y obtener recursos sin la

Recibido 03/03/2018
Aprobado 09/04/2018

Revista Publicando, 5. 14 (3). 2018, 571-582. ISSN 1390-9304

necesidad de realizar un gran cambio en su flujo de trabajo esta se puede considerar conforme a las expectativas de la comunidad. (Ochoa & Duval)

Consistencia lógica y coherencia: Los metadatos deben ser consistentes con los estándares y conceptos del dominio. La información contenida en los metadatos también debe tener coherencia interna, es decir, todos los campos de la instancia de metadato deben describir al mismo recurso de forma similar. (Ochoa & Duval)

Accesibilidad: Los metadatos que no pueden ser leídos o entendidos no poseen ningún valor. Si los metadatos van a ser utilizados por un GPS, el principal problema al que se puede enfrentar es a la accesibilidad física, por ejemplo, formatos incompatibles o enlaces rotos. Por otro lado, si los metadatos van a ser utilizados por los humanos, por ejemplo, la descripción, el mayor problema es la accesibilidad cognitiva (el metadato es muy difícil de entender). Estas dos dimensiones deben ser combinadas para estimar que tan difícil es acceder y comprender la información presente en los metadatos. (Ochoa & Duval)

Puntualidad: Los metadatos deben cambiar cada vez que el recurso es modificado. La puntualidad en un repositorio digital se relaciona principalmente con el grado en que una instancia de metadato sigue siendo actual. La actualidad de una instancia de metadatos podría medirse como lo útil que los metadatos permanecen con el paso del tiempo. (Ochoa & Duval)

Para el desarrollo de la aplicación fue seleccionado el parámetro de validación de metadatos completitud, debido a su facilidad de implementación y a que es ideal para recursos que aún no se encuentran dentro del repositorio. Los demás parámetros requieren el acceso a todos los recursos e instancias de metadatos de un repositorio, leer todo el contenido de un recurso o procedimientos de cálculo más complicados que pueden afectar la velocidad de respuesta de la propuesta de solución.

Resultado de la investigación

El resultado de la investigación es la aplicación web DEPXOS encargada de realizar el procesamiento en lote de documentos digitales para el sistema REPXOS 3.0. La misma permite a los administradores de REPXOS 3.0 la gestión de recursos digitales y la extracción de sus metadatos, que incluye su validación utilizando para esto el parámetro completitud, basado en pesos definidos previamente por el administrador.

Recibido 03/03/2018
Aprobado 09/04/2018

Revista Publicando, 5. 14 (3). 2018, 571-582. ISSN 1390-9304

Una vez extraídos los metadatos de los recursos, dichos recursos comienzan a ser tratados en el sistema como ítems (el documento digital y sus metadatos). Los ítems pueden ser eliminados completamente del sistema o editar sus metadatos de manera opcional. Una vez listos, los administradores pueden realizar el depósito de los mismos de manera individual o todos a la vez, siempre seleccionando previamente a qué colección realizar el depósito.

La comunicación con REPXOS 3.0 se realiza a través del estilo de arquitectura REST, de esta forma la autenticación en DEPXOS se realiza usando las credenciales propias de los administradores de REPXOS 3.0 y se garantiza que solo ellos puedan realizar el depósito.

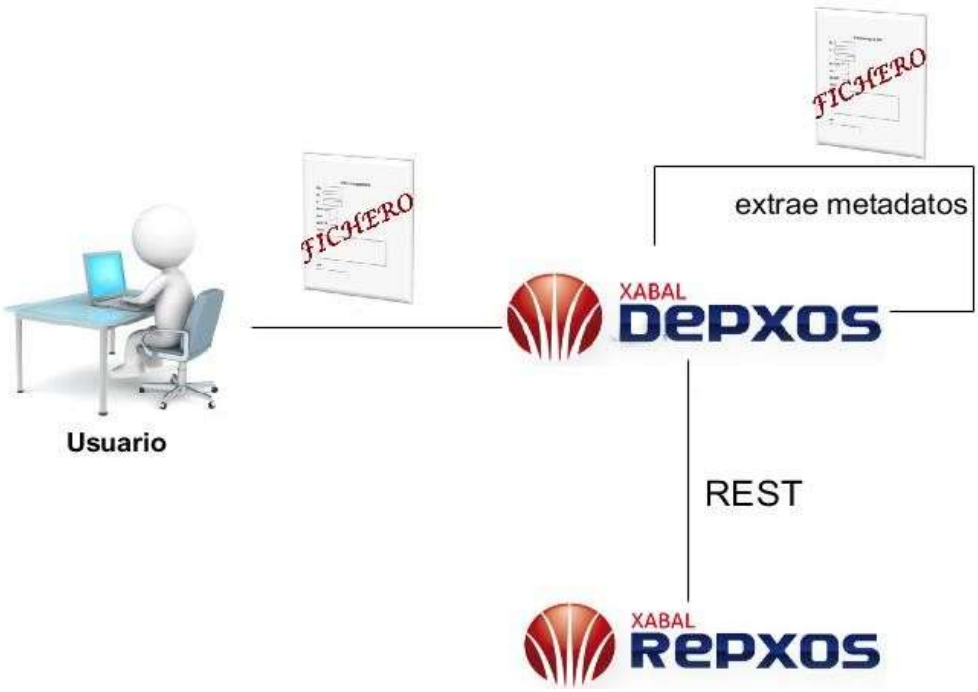

Figura 1. Flujo de trabajo en DEPXOS

4. CONCLUSIONES

Después de la investigación y posterior implementación se cumplió con el objetivo propuesto desarrollándose una herramienta que permite realizar el procesamiento en lote de documentos digitales para su posterior depósito en el sistema REPXOS 3.0. De manera general se obtuvieron los siguientes resultados:

Se caracterizaron las tendencias actuales a nivel mundial de la utilización de sistemas informáticos en la extracción automática de metadatos de documentos digitales

Recibido 03/03/2018
Aprobado 09/04/2018

Revista Publicando, 5. 14 (3). 2018, 571-582. ISSN 1390-9304

identificando ventajas y desventajas de las mismas, lo cual permitió seleccionar dos herramientas para usar en la propuesta de solución: GROBID y Apache Tika.

Se diseñó una solución que contribuyó al procesamiento en lote de los documentos digitales, que permite administrar dichos documentos, extraer sus metadatos y depositarlos en REPXOS 3.0.

DEPXOS fue utilizado en el Instituto de Geología y Paleontología y en la Fiscalía General de la República, escenarios ideales para implantar el sistema DEPXOS teniendo en cuenta la gran cantidad de documentos digitales que se generan, obteniéndose buenos resultados, siempre que los documentos a procesar cumplan con la estructura de publicación técnica o científica, o sus metadatos sean correctos. En el futuro se tiene pensado implementar otros parámetros de validación de metadatos contemplados en el marco de trabajo Bruce & Hillman además de añadirle un conjunto mayor de funcionalidades para mejorar la administración del sistema.

5. REFERENCIAS BIBLIOGRÁFICAS

Alvarez, M. A. (n.d.). *Manual de jQuery.*

Brito, N. (2009, junio 9). *Manual de desarrollo web con GRAILS.* Retrieved febrero 21, 2016, from http://www.manual-de-grails.es

CHRIS A. MATTMANN, J. L. (2012). *Tika in Action.* NY 11964: Manning Publications Co.

DSpace, E. d. (2015). *DSpace 5.x Documentation.* Retrieved from https://wiki.duraspace.org/display/DSDOC5x

Eleven Paths. (n.d.). Retrieved from https://www.elevenpaths.com/es/labstools/foca-2/

GROBID Documentation. (2016). Retrieved from http://grobid.readthedocs.org/en/latest/

Learn REST: A Tutorial. (2008 , febrero miércoles, 13).

Ochoa, X., & Duval, E. (n.d.). Automatic evaluation of metadata quality in digital repositories. International Journal on Digital .

Pérez, J. E. (2009). *Introducción a JavaScript* .

Pérez, J. E. (n.d.). *Introducción a AJAX.*

Recibido 03/03/2018
Aprobado 09/04/2018

Revista Publicando, 5. 14 (3). 2018, 571-582. ISSN 1390-9304

scholarlycommunication. (n.d.). Retrieved from http://scholarlycommunication.library.tamu.edu/repository-getting-started/policies/metadata-policy.html

Stuart Lewis, L. H.- W. (n.d.). *If SWORD is the answer, what is the question?*

The SWORD course - How SWORD Works. (n.d.).

Zealand, L. N. (2003). *Metadata Extraction Tool.* Retrieved febrero 11, 2016, from http://www.natlib.govt.nz/services/get-advice/digital- libraries/metadata-extraction-tool

Recibido 03/03/2018
Aprobado 09/04/2018