REVISTA COLOMBIANA DE COMPUTACIÓN
Volumen 15, número 2
Págs. 135 - 160

Extracción Automática de Metadatos como Soporte para el Autoarchivo de Objetos Digitales en Repositorios

Ana Casali[*†], Cristina Bender[*‡], Claudia Deco[*‡], Santiago Fontanarrosa[*]

Fecha de recibido: 28/062014 Fecha de Aprobación: 09/09/2014

Resumen

En este trabajo se propone facilitar al usuario el autoarchivo de sus objetos digitales educativos en un repositorio institucional. Para esto, se modifica el flujo de carga estándar de la plataforma DSpace, proponiendo un nuevo flujo para el depósito de objetos de modo que pueda integrarse en este proceso un extractor de metadatos. Se presenta una arquitectura abierta de un módulo extractor automático de algunos metadatos de los documentos. Estos metadatos extraídos automáticamente son luego validados por el usuario en el proceso de descripción del objeto. Para diseñar el extractor se analizaron distintas herramientas de extracción y se optó por la combinación que arrojó mejores resultados. Este módulo se ha diseñado de modo de poder integrar otras herramientas extractoras. Se ha desarrollado un prototipo en JAVA de este asistente y se ha experimentado sobre dos corpus de documentos, uno en idioma Inglés y otro en idioma Español. En este trabajo, se presentan resultados de la extracción automática de los metadatos Palabras clave, Título y Autores, en documentos en ambos idiomas, los cuales resultan promisorios. Mediante este asistente se espera ayudar al usuario en el proceso de carga de objetos digitales educativos disminuyendo así su trabajo, y mejorando la cantidad y la calidad de los metadatos cargados.

Palabras Clave: *Extracción Automática de Metadatos, Objetos Digitales Educativos, Repositorios, Flujo de Carga*.

Abstract

This paper aims to facilitate users the self-archiving of his/her digital learning objects in an institutional repository. For this, the standard load flow of DSpace

[*]Departamento de Sistemas e Informática, Facultad de Ciencias Exactas, Ingeniería y Agrimensura, Universidad Nacional de Rosario. Av. Pellegrini 250, 2000 Rosario, Argentina.

[*†]Centro Internacional Franco Argentino de Ciencias de la Información y de Sistemas CIFASIS. 27 de Febrero 210bis, 2000 Rosario, Argentina.

[*‡]Departamento de Investigación Institucional, Facultad de Química e Ingeniería Rosario, Pontificia Universidad Católica Argentina, Av. Pellegrini 3314, 2000 Rosario, Argentina.

platform changes, proposing a new flow for the object repository so that it can be integrated a metadata extractor in this process. An open architecture for automatic metadata extraction is presented. These metadata are automatically extracted and then is validated by the user in the process of object description. In order to design the extractor, different extraction tools were analyzed and we chose the combination that yielded the best results. This module is designed so that other extraction tools may be integrated. We have developed a prototype in JAVA and an experimentation was performed on two corpus of documents, one in English and the other in Spanish. In this paper, results of the automatic extraction of Keywords, Title, and Authors metadata are presented. Using this assistant is expected to help users in the process of uploading digital learning objects, decreasing their work, and improving the quantity and quality of the metadata loaded.

Keywords: *Automatic Metadata Extraction, Digital Learning Objects, Repositories, Load Flow.*

1. Introducción

En los últimos años, el desarrollo de repositorios institucionales de acceso abierto ha sido un tema prioritario en las políticas de educación, ciencia y técnica de muchos países y en particular, la creación de repositorios de objetos digitales educativos en las universidades públicas de Argentina, es una prioridad en el marco de las políticas de nuestro país. El objetivo de estos repositorios es viabilizar de una forma eficiente el almacenamiento, clasificación, búsqueda y reutilización de recursos educativos. En este sentido, se proponen repositorios universitarios integradores de Objetos Digitales Educativos donde una publicación científica o una obra de arte pueden ser también consideradas objetos de aprendizaje.

Un objeto de aprendizaje es "cualquier recurso digital que puede ser reutilizado para la enseñanza" [1]. Estos pueden ser usados por un estudiante que quiere aprender un determinado tema o por un profesor que quiere preparar algún material para su clase. Los usuarios pueden recuperar estos objetos por medio de búsquedas en repositorios Web. Los Repositorios Institucionales almacenan la producción docente, científica y de extensión, y permiten una búsqueda más acotada para la recuperación y reutilización de estos recursos digitales.

Estos objetos se almacenan utilizando metadatos descriptivos que proporcionan información adicional sobre el mismo. La información almacenada en estos metadatos es fundamental para la mejor recuperación de los mismos y se vuelven un aspecto clave en el rendimiento y calidad de los objetos retornados. Existen distintos estándares de metadatos tales como Dublin Core (dublincore.org) y

IEEE LOM (www.ieee.org), que utilizan distintas categorías no solo para describir el contenido del objeto (título, autor, palabras claves, idioma, etc.) sino también, como en el caso de LOM, permiten describir aspectos educacionales de los mismos (nivel educativo, complejidad, etc.). Sin embargo, en la mayoría de los casos, la información cargada en estos metadatos en los distintos repositorios, es de baja calidad o incompleta ([2], [3]). Esto se debe a que la carga de metadatos, es una tarea que suele ser tediosa, que consume tiempo y muchas veces las personas encargadas de llenar los mismos, deciden no hacerlo.

La Universidad Nacional de Rosario ha creado en los últimos años un Repositorio Hipermedial institucional, denominado RepHip (rephip.unr.edu.ar) cuyo objetivo es almacenar toda la producción académica, científica y de extensión de la misma. Este repositorio está implementado en la plataforma DSpace (dspace.org). Esta plataforma se ha adoptado para implementar el Sistema Nacional de Repositorios Digitales (SNRD) (repositorios.mincyt.gob.ar) de Argentina. En este repositorio se trabaja por comunidades y cada una de ellas es la encargada del autoarchivo de sus documentos.

Para facilitar la carga de objetos digitales educativos en el repositorio se propone modificar el flujo de carga estándar de la plataforma DSpace y diseñar un asistente de carga para la extracción automática de algunos metadatos, extracción que deberá realizarse en tiempo real. De esta forma se ayudará al usuario en este proceso de carga con el objetivo de mejorar la cantidad y la calidad de los metadatos cargados.

Se ha propuesto un nuevo flujo de carga para el depósito de objetos y una arquitectura para incorporar la extracción automática de metadatos de los objetos a cargar en repositorios desarrollados sobre DSpace. Una propuesta preliminar de esta arquitectura y una comparación de extractores sobre un corpus de documentos del Repositorio RepHip se ha presentado en [4]. Estos resultados preliminares han mostrado ser promisorios y permitieron tomar decisiones para el diseño del módulo extractor. A partir de esta arquitectura, en este trabajo se presenta una nueva experimentación utilizando el prototipo implementado y eligiendo otro corpus de otro repositorio para evaluar la performance del sistema extractor en otro contexto. Además, se evalúa en forma separada los documentos según el idioma en que están escritos, inglés o español, midiendo la precisión y a diferencia del trabajo presentado en [4] se mide también la cobertura obtenida en cada uno de los metadatos extraídos. Estas métricas han sido tomadas de la recuperación de información y han sido adaptadas a los casos bordes siguiendo la propuesta utilizada ([5], [6]), donde se presenta una arquitectura para asistir la recopilación de documentos de texto plausibles de ser cargados en Repositorios Institucionales, que incluye un módulo extractor de

metadatos para catalogar al documento y obtener datos de filiación y contacto de los autores.

Este trabajo está estructurado de la siguiente forma: en la Sección 2, se describe el flujo de carga de la plataforma DSpace y las mejoras planteadas en dicho flujo para la incorporación de un módulo de extracción de metadatos; se presentan algunos conceptos referidos a la Extracción de Información y las métricas utilizadas para su evaluación; se presenta, además la problemática de extracción de metadatos y el análisis de algunas herramientas de extracción automática; y se describe la arquitectura propuesta del módulo extractor de metadatos. En la Sección 3, se describe el diseño de la experimentación y se analizan los resultados de la misma. Finalmente en la Sección 4, se presentan algunas conclusiones.

2. Metodología

2.1. Población de Repositorios: el Caso de Dspace

La plataforma DSpace (dspace.org) se ha adoptado para implementar el Sistema Nacional de Repositorios Digitales (repositorios.mincyt.gob.ar) de Argentina. Esta plataforma utiliza Colecciones para agrupar los documentos, las cuales son manejadas por un Administrador de la Comunidad/Subcomunidad. Para cargar un documento el usuario selecciona una comunidad, y en base a esa selección se determinan los tipos de objetos digitales que el usuario podrá elegir. Por ejemplo, la comunidad *Departamento de Ciencias de la Computación* tiene asociadas las colecciones *Tesinas*, *Artículos*, *Comunicaciones a Congresos* y *Materiales Educativos*.

El proceso de carga por defecto de objetos en repositorios gestionados por DSpace se divide en una serie de pasos, como se muestra en la Fig. 1. En estos pasos se eligen las colecciones en las que se va a realizar el depósito, se describe el objeto mediante metadatos, se suben los archivos que lo componen, se acepta la licencia del repositorio y se hace una revisión de los metadatos previa a que se complete el depósito. También, se puede agregar un paso extra que permite elegir una licencia Creative Commons (creativecommons.org) para el objeto a depositar antes de la revisión. Cada una de estas tareas debe realizarse en el orden mencionado. La descripción del objeto mediante metadatos se realiza en tres etapas: una de preguntas iniciales, la segunda para cargar los metadatos obligatorios y la tercera para cargar metadatos opcionales del objeto.

A partir de un análisis de usabilidad de DSpace, realizado por usuarios docentes-investigadores, con habilidades informáticas de distinto

grado, se encontraron los siguientes problemas en el proceso de depósito:

- *La tarea de carga de metadatos es tediosa:* la descripción con metadatos del objeto a depositar es un gran cuello de botella en el depósito ya que mucha de la información requerida no está al alcance inmediato de la persona que realiza el depósito y no se completa.
- *La interfaz del flujo de depósito no es clara:* existen problemas de interfaz que hacen confuso el depósito. Algunos de los problemas encontrados son: no es claro cómo comenzar una nueva carga, a pesar de ser una de las funcionalidades principales del repositorio; la elección de colecciones en donde éste se va a realizar es confusa ya que no especifica a qué comunidad pertenecen; los metadatos obligatorios no están diferenciados claramente de los metadatos opcionales; los distintos pasos de descripción tienen el mismo nombre y algunos botones del flujo de carga son ambiguos.
- Los *pasos para la descripción del objeto no son customizables según los distintos tipos de colecciones*: si bien DSpace permite la configuración de los formularios de descripción a partir de un archivo de configuración simple, esta debe hacerla un administrador para cada colección; además, no permite agrupar colecciones en distintos tipos y elaborar depósitos configurados para cada tipo.

Fig. 1. Flujo de carga de la plataforma Dspace.

2.1.1 Mejoras Planteadas

Para solucionar estos problemas, en [7] se propone reestructurar el flujo de carga reordenando y modificando los pasos de depósito, modificar la interfaz e incorporar un módulo de extracción de metadatos de los archivos depositados. Las modificaciones principales al flujo de carga se muestran en la Fig. 2 y los pasos propuestos se describen a continuación.

1. *Elección de la colección:* el nuevo paso muestra la estructura completa de comunidades y colecciones en las que el usuario tiene permisos de realizar el depósito.
2. *Aceptación de la licencia institucional:* este paso se reubicó al principio del proceso ya que en el caso de que esta no se acepte, el depósito se cancela y los otros pasos no son necesarios.
3. *Carga de los archivos asociados al objeto:* esta etapa también se reubicó y se realiza antes de completar los formularios de descripción, de manera que algunos metadatos puedan extraerse automáticamente con ayuda del asistente de extracción de metadatos.
4. *Reordenamiento de la descripción mediante dos pasos bien diferenciados:* uno para la carga de metadatos obligatorios y otro para metadatos opcionales. Estos metadatos cambian según el tipo de objeto que se está depositando y es inferido por la colección que se eligió. Se propone colaborar con el usuario en la carga de los metadatos, mediante el agregado de un módulo externo que extraiga estos metadatos del archivo a cargar en el repositorio y que se muestren al usuario para su validación en los campos correspondientes del formulario.

Fig. 2. Flujo de Carga propuesto.

2.2. Extracción de Información

El objetivo principal de los sistemas de Extracción de Información, es localizar información a partir de documentos de texto en lenguaje natural, produciendo como salida del sistema, un formato tabular estructurado, en un formato fijo de datos sin ambigüedad, que pueden ser resumidos y presentados de manera uniforme [5]. La representación uniforme de los datos y sus relaciones resulta conveniente para la inspección y comparación de hechos que pueden ser visualizados con mayor facilidad en una tabla. Además, contando con una representación uniforme y estructurada de un documento, éste puede analizarse con herramientas automáticas tales como técnicas de minería de datos para el descubrimiento de patrones y posterior interpretación de los mismos. Así, una de las principales tareas de los sistemas de extracción de información consiste en encontrar la información relevante del documento de entrada, donde, por supuesto, la relevancia se define de acuerdo con la información que se espera encontrar.

La Extracción de Información refiere a una tarea diferente de la Recuperación de Información. Esta última, se enmarca en un campo mucho más antiguo y maduro en el que la principal atención está puesta en la selección de un subconjunto de documentos dentro de un conjunto mucho más grande a través de una consulta. El usuario de un sistema de recuperación de información, luego de una consulta, debe examinar el/los documentos de la salida para la consiguiente extracción de información. El contraste entre los objetivos de las tareas de recuperación de información con la de extracción de información, puede entonces resumirse en: un sistema de recuperación de información recupera los documentos relevantes dentro de una colección más grande, mientras que un sistema de extracción de información extrae información relevante dentro de uno o más documentos. Por lo tanto, ambas técnicas son complementarias y utilizadas en combinación pueden resultar en herramientas poderosas de procesamiento de texto. Ambas áreas difieren además de en sus objetivos, en las técnicas de cómputo utilizadas. Estas diferencias se deben en parte a los objetivos inherentes de las mismas y en parte a la historia de cada una de las áreas. Gran parte del trabajo que ha emergido en la extracción de información proviene de sistemas basados en reglas, computación lingüística y procesamiento de lenguaje natural, mientras que en el campo de la recuperación de información, han influido áreas como teoría de la información, probabilidad y estadística.

2.2.1 Métricas de Evaluación

Una de las principales tareas de la extracción de información consiste en encontrar la información relevante del documento de entrada, donde la relevancia se define según la información que se espera encontrar ([8],

[9]). La necesidad de métricas de evaluación en los problemas de extracción de información apareció rápidamente en las conferencias sobre la problemática [6]. El punto de partida para el desarrollo de las mismas fueron las métricas Cobertura y Precisión de la Recuperación de Información ([10], [11]). Aunque las medidas para la extracción no son las mismas, los nombres se han mantenido. Las alteraciones a las mismas permiten reportar posibles errores producidos en el intento de un algoritmo de extracción de generar información.

En la extracción de información, se definen la Cobertura y la Precisión de la forma siguiente:

$$C \quad \frac{\textit{Cantidad respuestas correctas}}{\textit{Cantidad total posible de respuestas correctas}} \qquad (1)$$

$$P \quad \frac{\textit{Cantidad respuestas correctas}}{\textit{Cantidad de respuestas producidas}} \qquad (2)$$

Esto es, la cobertura (*C*) es la razón entre la cantidad de respuestas correctas respecto a la cantidad total posible de respuestas correctas; y la precisión (*P*) es la razón entre la cantidad de respuestas correctas respecto a la cantidad de respuestas producidas. Ambas métricas, cobertura y precisión toman valores siempre en el intervalo [0,1] y el valor óptimo para cada una es 1.

Para evaluar el desempeño de la propuesta presentada en este trabajo, siguiendo a [5] se aplicaron convenciones usuales en el área para los casos bordes:

- Si la Cantidad total de respuestas correctas es nula (lo que indica que en la ecuación (1) el denominador es nulo): se asigna C=1 en el caso de que el extractor no produzca respuestas y C=0 en caso contrario. De esta forma, se premia o penaliza el hecho de haber producido o no datos espurios ante la ausencia de datos.
- Si la Cantidad de respuestas producidas es nula (lo que indica que en la ecuación (2) el denominador es nulo): se asigna un valor no numérico (N/A: No Aplicable) a la precisión, considerando que carece de importancia medir este valor en esta situación.

De esta forma utilizando ambas medidas, se diferencia la efectividad de un extractor en los resultados producidos y la cobertura sobre el total de datos a extraer. Además, se penaliza en la cobertura, la falsa extracción de datos (falsos positivos).

2.3 Extracción de Metadatos

La problemática de extracción de metadatos en recursos educativos es un problema abierto y difícil de abordar. Esto se debe a la diversidad de tipos de recursos, distintos formatos de archivos utilizados y falta o diversidad de estructura en los mismos. Este problema ha sido parcialmente abordado en [12] donde se analizaron algunos sistemas dedicados a la extracción automática de metadatos educativos de objetos de aprendizaje, que aunque son relevantes, algunas no están implementadas o no están disponibles como herramientas libres.

En este trabajo, se plantea focalizar en la extracción de metadatos en archivos en formato de texto. Para esto, se analizaron herramientas extractoras de metadatos generales tales como el título, los autores, las palabras claves, el resumen y el idioma. En un primer análisis se seleccionaron las siguientes herramientas: KEA, Mr. Dlib, Alchemy y ParsCit.

KEA Automatic Keyphrase Extraction (www.nzdl.org/Kea/indexold.html), es la implementación en JAVA del algoritmo KEA [13]. La herramienta extrae automáticamente frases claves del texto completo a partir del documento a analizar. El conjunto de todas las frases seleccionadas en un documento se identifican utilizando procesamiento léxico rudimentario. Utiliza técnicas de machine-learning para generar un clasificador que determina qué frases candidatas deben ser asignadas como frases clave. Esta herramienta puede ser utilizada en forma local y se necesita una fase previa de entrenamiento.

Mr. DLib (www.mr-dlib.org/) es una biblioteca digital que proporciona acceso a varios millones de artículos de texto completo y sus metadatos en formato XML y JSON a través de un servicio web RESTful. En su etapa beta de desarrollo, sus funcionalidades son utilizadas por terceros y permite extraer Título y Autores [14].

AlchemyAPI (www.alchemyapi.com) es una plataforma de minería de texto la cual proporciona un conjunto de herramientas que permiten el análisis semántico utilizando técnicas de procesamiento de lenguaje natural. Provee un conjunto de servicios que permiten analizar de forma automática documentos de texto. La herramienta expone varios servicios a partir de su RESTful API (www.alchemyapi.com/ api/calling.html), entre los que se encuentran: extracción de Autor, Entidades, Palabras claves, categorización del Contenido e identificación del Idioma. En su versión gratuita, el servicio presenta una limitación de 1000 consultas diarias y un límite por consulta de 150 kbs.

ParsCit (wing.comp.nus.edu.sg/parsCit/) es una aplicación de código abierto que realiza dos tareas: el análisis sintáctico de cadenas de

referencia, también llamado análisis de citas o extracción de citas, y el análisis de la estructura lógica de documentos científicos. Estas tareas las realiza a partir de un archivo de texto plano utilizando procedimientos de aprendizaje automático supervisado que usan campos aleatorios condicionales (CRF) como mecanismo de aprendizaje. Incluye utilidades para ejecutarse como un servicio Web o como una aplicación independiente [15].

Para evaluar los metadatos generados por las distintas herramientas extractoras se ha comparado la lista generada con la extracción de estos metadatos realizada manualmente a partir de la revisión de cada documento tratado. En nuestro trabajo, en una primera instancia se ha considerado la evaluación de la Precisión, teniendo en cuenta solo los conceptos relevantes recuperados.

Para elegir los componentes a utilizar en el módulo extractor de metadatos, se siguieron los resultados obtenidos previamente en [4]. En dicho trabajo se realizaron distintas experimentaciones con las herramientas mencionadas, sobre un corpus de 760 documentos del repositorio RepHip seleccionados según los siguientes criterios: diversidad temática, de colecciones y de tipo de archivo. Las pruebas que se realizaron buscaban evaluar los resultados obtenidos al realizar la extracción de títulos, autores, palabras claves e idioma así como el tiempo de respuesta, ya que es una condición que los resultados se obtengan en tiempo real. Se analizaron los resultados obtenidos por Kea, Alchemy y Mr. DLib, respecto a los distintos metadatos que pueden ser extraídos por cada uno ellos: Mr. DLib para título y autores, KEA para palabras claves y Alchemy para la extracción de título, palabras claves e idioma.

En el caso de Mr. Dlib, se realizaron varias pruebas: enviando las primeras, 1, 2 o 4 páginas del archivo PDF al servidor de Mr. Dlib para la extracción. El objetivo fue analizar si se obtienen extracciones más precisas y si los tiempos de procesamiento eran aceptables. Los resultados son similares, salvo que los tiempos de procesamiento por artículo en el caso de 4 páginas se duplican (los promedios van de 5,6 seg a 11,2 seg). Si bien los resultados de extracción de títulos y autores de los documentos, utilizando Mr. DLib fueron parcialmente satisfactorios (precisión 65 % para los títulos y 40 % para autores) se lo descartó para este primer prototipo ya que esta herramienta estaba en una etapa muy temprana de desarrollo y su accesibilidad no estaba garantizada.

Respecto a KEA, se lo configuró para sugerir 5 palabras clave por documento, y se compararon las Palabras Clave extraídas vs Palabras Claves previamente ingresadas. El promedio de coincidencias fue de un 60% con un tiempo de extracción promedio por documento de 1,79 segundos, basado en el corpus de documentos con tamaños de hasta 800 KBytes.

Las pruebas para evaluar AlchemyAPI consistieron en enviar los archivos al servidor que provee AlchemyAPI, a fin de obtener la siguiente metadata: Título, Palabras Clave e Idioma. En la extracción del idioma del documento, se obtuvo un 76% de asignaciones correctas. Respecto a las palabras claves, Alchemy retorna un ranking de relevancia. Para su evaluación, se consideraron únicamente las primeras 5, y se compararon los resultados obtenidos con las cargadas por el autor del documento. Se obtuvo un 56% de resultados correctos, donde se recuperaron todas o algunas de las palabras claves cargadas por el autor. Respecto al título el 67% de las asignaciones fueron correctas.

Se observa que los resultados encontrados con KEA y Alchemy respecto a palabras claves son similares en precisión y que los resultados obtenidos con Mr. DLib y Alchemy para título y autores, también lo son. Como la herramienta Alchemy también permite obtener el idioma, se plantea utilizarla en el prototipo combinándola con otra herramienta para el preprocesamiento de documentos como ParsCit (wing.comp.nus.edu.sg/parsCit/) para mejorar los resultados. ParsCit permite dar estructura al documento y genera en un documento XML en el cual intenta identificar: Título, Autor, Resumen y Palabras claves. Esta información se concatena en un nuevo archivo, el cual se utiliza para subir al servidor de AlChemyAPI en lugar del archivo original.

En una segunda etapa de prueba, considerando el mismo conjunto de documentos, se agregó un paso más en el proceso de extracción de palabras clave, ejecutando primero la herramienta ParsCit y luego Alchemy. Se obtuvo en este caso, el 70% de resultados correctos, considerando como correctos aquellos donde se recuperaron todas o algunas de las palabras claves cargadas por el autor. Entonces, a partir de esta combinación de herramientas, se logró incrementar sustancialmente la calidad de las palabras claves retornadas pasando de un 56% de resultados correctos obtenidos con Alchemy a un 70% resultante con ParsCit+Alchemy. Los resultados obtenidos se resumen en la Tabla 1.

Al analizar los resultados erróneos, se observó que esto se debió a que algunos documentos no reportaron datos y/o no pudieron ser analizados. La imposibilidad de dicho análisis se debió a alguno de los siguientes problemas: El archivo presentaba un formato que no permitía extraer y transformar el contenido del mismo a texto plano (por ejemplo. archivos ppt convertidos a pdf) lo cual hacía que la calidad de la extracción fuera baja y/o incluyera la mayoría de metadatos propios del formato, los cuales no aportan valor real al extractor; la cantidad de texto plano extraído del documento original no era suficiente para que el servicio de Alchemy pudiera generar una respuesta; o el formato del documento original no era soportado por la herramienta que debe transformarlo a texto plano.

Herramienta	Título	Autor	Palabras Claves	Idioma
Kea			60 %	
Mr. DLib	65 %	40 %		
AlchemyAPI	67 %	40 %	56 %	76 %
AlchemyAPI+ParsCit	67 %	68 %	70 %	76 %

Tabla 1. Precisión obtenida en la extracción de metadatos con las herramientas analizadas.

A partir de los resultados mostrados en la tabla anterior, se decidió utilizar ParsCit para dar estructura al documento, además de permitir la extracción de título y autor, y a AlchemyAPI para la generación de palabras claves.

2.4 Arquitectura Propuesta para el Módulo Extractor de Metadatos

Con el fin de ayudar al usuario en el proceso de autoarchivo de documentos se ha diseñado un módulo de extracción que permite la extracción automática de metadatos en documentos que estén, o puedan convertirse, a formato de texto. El tiempo de respuesta del módulo es relevante, ya que es una condición que los resultados se obtengan en tiempo real. Este módulo se puede implementar en DSpace a partir de las modificaciones en el flujo de carga propuestas. Del análisis de herramientas para la extracción de metadatos descrito en la sección anterior se ha decidido a partir de un archivo en formato texto, utilizar ParsCit para dar estructura al documento generando un documento XML, el cual se utiliza para enviar al servidor de AlChem y API para la extracción de los demás metadatos. La Fig. 3 muestra la arquitectura propuesta y cómo ésta interacciona con el flujo de carga de Dspace.

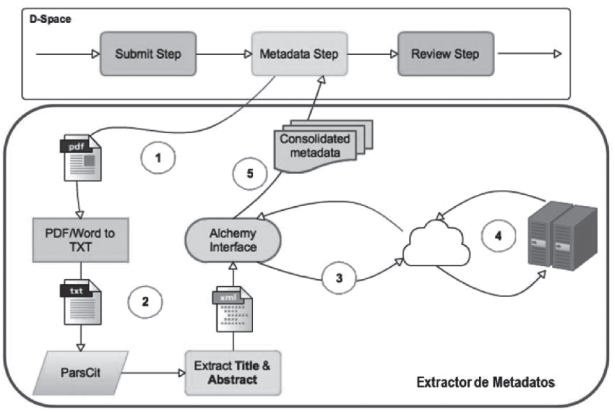

Fig. 3. Arquitectura propuesta y su interacción con Dspace

Una vez que el usuario ha realizado los primeros pasos del flujo de carga (resumidos en Submit Step en la figura), es decir la selección de la colección en la cual quiere depositar el documento, la aceptación de la licencia institucional y la carga de los archivos asociados al objeto se envía al Extractor el documento principal que se utilizará para la extracción de metadatos.

El módulo Extractor de Metadatos interactúa con el Metadata Step que está desglosado en los pasos Descripción de Metadatos Obligatorios y Descripción de Metadatos Opcionales (como se muestra en la Figura 2). El Metadata Step invoca al módulo Extractor de Metadatos enviando el contenido del archivo que se está cargando en el repositorio. Este módulo obtiene los metadatos resultantes del proceso de extracción y es responsable de parsear los resultados y presentarlos de forma adecuada para que DSpace los muestre al usuario.

El Extractor de Metadatos tiene una arquitectura interna de *pipe-line*, facilitando así la posibilidad de agregar nuevos módulos al proceso, y está compuesto de los siguientes submódulos.

1. *Conversión a archivo de texto:* este submódulo toma como entrada el archivo que cargó el usuario (en formato pdf, doc, ppt, etc.) y extrae el contenido en formato de texto (.txt). DSpace almacena dicho archivo en su representación de bytestream.
2. *Estructuración del documento:* recibe como entrada el archivo en texto plano, organiza y da estructura al mismo generando un archivo XML en el cual se puede identificar título, autor, resumen y palabras claves. Para ello se utiliza la herramienta ParsCit (wing.comp.nus.edu.sg/parsCit/).

3 y 4. *Extracción complementaria:* recibe como entrada el archivo en formato XML con la estructura y utiliza dicha estructura para enviar al servicio Alchemy, solamente el texto con mayor probabilidad de tener información que resulte relevante. Alchemy envía la información al servidor extrayendo además el idioma y las palabras claves. El servidor retorna la respuesta en formato JSON.

5. *Integración de resultados:* se consolida la respuesta del servidor y se envía a DSpace un archivo XML con todos los metadatos extraídos automáticamente. Este módulo es responsable de integrar la respuesta y transformarla en un formato reconocible por el proceso de carga de Dspace.

Una vez finalizada la generación de metadatos, el proceso de carga de documentos en DSpace continúa de forma normal. Luego de que el usuario valide y/o modifique estos metadatos obtenidos automáticamente y complete la descripción de los faltantes, se pasa a la verificación por parte del usuario de todos los datos cargados (Review

Step) finalizando así el depósito del objeto en el repositorio.

Para validar la arquitectura propuesta, se implementó un prototipo de este asistente desarrollado en código Java. Las interfaces consisten en tres módulos que componen el generador de metadatos, forman una unidad única, la cual tiene una interfaz de entrada y una de salida. Para esta implementación se ha definido un metamódulo que agrupa los distintos componentes y se encuentra desarrollado en Python. Los parámetros de entrada son un camino a un archivo donde se encuentra el documento y la colección a la que pertenece dicho archivo según está definido en el contexto de DSpace. La respuesta que el script devuelve es un JSON donde se incluye el título, autores, idioma y palabras claves.

3. Resultados

3.1 Diseño de la Experimentación

Para la evaluación del prototipo, se realizó una experimentación con documentos extraídos de otro repositorio a fin de evaluar la generalidad de la propuesta. Para ello se generaron dos conjuntos de 100 documentos cada uno, extraídos del repositorio e-LIS (eprints.rclis.org). Los documentos utilizados, se eligieron con los siguientes criterios: que estuvieran en formato PDF y que correspondan a la categoría de tesis o de artículo. Para cada uno de los documentos elegidos, se procedió a obtener manualmente el título, los autores y el conjunto de palabras claves que los usuarios precargaron, a fin de poder utilizarlos como casos de referencia para poder evaluar los resultados generados. El primer conjunto contiene documentos en idioma inglés y el segundo contiene documentos en español.

El objetivo de la experimentación es evaluar el prototipo desarrollado y así validar la arquitectura propuesta. Además, la evaluación con los dos grupos de documentos en distintos idiomas permite contrastar la precisión y la cobertura logradas según el idioma del documento.

Dado que la carga y generación de metadata a través de la plataforma de DSpace requiere la intervención humana, y a los fines de agilizar el proceso de evaluación, se procedió a automatizar las pruebas mediante el uso de scripts que realizan las tareas simulando la interacción del usuario. Las funcionalidades que se automatizaron son las siguientes:

– *Descarga de documentos*: a partir del listado de documentos obtenidos por los criterios de búsqueda en el repositorio, se procede a descargar cada uno de ellos. Los archivos son descargados a disco

en su formato original (PDF) y se ingresa su registro correspondiente en la base de datos local de referencia. Adicionalmente, se le asocia la siguiente información: título del documento, autores y palabras claves que fueron cargadas por el autor del documento.

- *Extracción del texto plano*: una vez que se cuenta con el conjunto de documentos de prueba, se procede a extraer el texto en formato plano. Esto es necesario dado que el módulo Extractor de Metadatos acepta como parámetro de entrada el archivo en dicho formato. Para esta tarea se utiliza la herramienta pdf2text (en.wikipedia.org/wiki/Pdftotext).
- *Generación de los metadatos*: en esta instancia, se procede a invocar al módulo de generación de metadatos, simulando que se estuviera haciendo dentro del proceso de carga normal de D-Space. El resultado es guardado a disco, para su posterior procesamiento.
- *Procesamiento de resultados*: todos los resultados son guardados de forma local utilizando el motor de base de datos MySQL (www.mysql.com). Por cada archivo a evaluar, se registran todos los campos extraídos por la herramienta, para su posterior evaluación.

A continuación se presentan los análisis efectuados para los metadatos Palabras clave, Título y Autores.

3.1.1 Metadato: Palabras Claves

Durante la primera etapa de experimentación, se ejecutaron casos de prueba sobre un conjunto reducido de documentos, a fin de evaluar la efectividad de las validaciones. Inicialmente se observó que *ParsCit* fallaba en detectar las palabras clave en los documentos, aun cuando estas contaban con su propia sección y se especificaban de tal forma por ejemplo, con el término Keywords o la frase Palabras Claves en un párrafo separado. Analizando en mayor detalle los documentos en los cuales se observaba este comportamiento, se pudo detectar que dada la particularidad de algunos documentos, los cuales no seguían el formato tradicional de tesis o de artículo, estas secciones eran agrupadas erróneamente en otras áreas como por ejemplo, en el resumen. En la mayoría de los casos, esto se debía a que el formato que seguían estos documentos no respetaba la misma segmentación, o en algunos casos el formato de múltiples columnas tendía a romper el flujo natural del texto, con lo cual el algoritmo fallaba al dar estructura. Debido a esto, se procedió a refinar el proceso de generación de metadata. Para el refinamiento, se realiza una búsqueda adicional en los resultados, consistente en efectuar dentro

del resultado de *ParsCit*, la detección de palabras no significativas específicas, tales como los términos *Keywords* o *Palabras Clave*. De esta forma se evita detectar palabras clave en secciones erróneas y se mejoran los ítems detectados en la primera etapa. Solucionada esta dificultad, se procedió a analizar los resultados, asignando valores numéricos a las extracciones, como se describe a continuación.

En las Tablas 2 y 3 se muestra la asignación de valores, a modo de ejemplo, para algunos documentos y los valores de cobertura y precisión correspondientes a cada uno.

ID-Archivo	Respuestas Válidas Posibles (a)	Respuestas Correctas Generadas (b)	Respuestas Totales Generadas (c)	Cobertura ((b) / (a)) en %	Precisión ((b) / (c)) en %
DE-1	2	1	4	50%	25%
DE-2	6	5	10	83%	50%
DE-3	12	10	12	83%	83%
DE-4	8	7	8	87%	87%
DE-5	9	9	10	100%	90%
DE-6	19	19	20	100%	95%
DE-7	10	10	10	100%	100%

Tabla 2. Ejemplo de asignación de valores para Palabras clave en documentos en español.

ID-Archivo	Respuestas Válidas Posibles (a)	Respuestas Correctas Generadas (b)	Respuestas Totales Generadas (c)	Cobertura ((b) / (a)) en %	Precisión ((b) / (c)) en %
DI-1	3	1	10	33%	10%
DI-2	7	7	7	100%	100%
DI-3	15	15	16	100%	94%
DI-4	16	15	16	94%	94%
DI-5	4	0	1	0	N/A
DI-6	8	0	2	0	N/A
DI-7	6	3	3	50%	100%

Tabla 3. Ejemplo de asignación de valores para Palabras claves en documentos en inglés.

En estas tablas, la primera columna contiene la identificación del archivo; la segunda columna contiene la cantidad de respuestas válidas posibles, esto es, las respuestas que debería producir el extractor y que se obtuvieron por medio de la inspección manual del documento; la

tercera columna presenta la cantidad de respuestas correctas generadas por el extractor, las cuales fueron determinadas a partir de la comparación manual de las respuestas generadas por el extractor contra las obtenidas por inspección manual de los documentos; la cuarta columna muestra la cantidad total de respuestas generadas por el extractor; la quinta columna contiene la cobertura y la sexta columna la precisión alcanzadas por el extractor para ese documento, según lo discutido en la Sección 2.2. La Tabla 2 presenta los documentos en idioma español y la Tabla 3 los documentos de ejemplo en idioma inglés.

En la Tabla 3, se muestran dos documentos (identificados con DI-5 y DI-6) de los cuales no se generaron respuestas correctas, por lo que, siguiendo con el criterio presentado en la Sección 2.2 se asigna un valor 0 (cero) para la Cobertura y no se evalúa la Precisión (N/A).

3.1.2 Metadato: Título

Para este metadato, se analizaron ambos conjuntos de 100 documentos, y a modo de ejemplo se tomaron los mismos subconjuntos de documentos tratados en la sección anterior, para mostrar la asignación de valores utilizada. En las tablas siguientes (Tabla 4 y Tabla 5) se muestra esta asignación para los Títulos de los documentos.

ID-Archivo	Titulo Real	Titulo Encontrado	Grado de coincidencia	Cobertura	Precisión
DE-1	Tecnología digital en Bibliotecas en Paraguay	Tecnología digital en Bibliotecas en Paraguay	2	100%	100%
DE-2		---	0	0	N/A
DE-3		---	0	0	N/A
DE-4	Media Literacy for Older People facing the Digital Divide: The e-Inclusion Programmes Design	Media Literacy for Older People facing the Digital Divide: The e-Inclusion Programmes Design	2	100%	100%
DE-5		---	0	0	N/A
DE-6		---	0	0	N/A
DE-7	Trayectoria tecnológica Web y el orden digital en Latinoamérica: reflexiones históricas desde Brasil.	Trayectoria tecnológica Web y el orden digital en Latinoamérica: reflexiones históricas desde Brasil.	2	100%	100%

Tabla 4. Ejemplo de asignación de valores para Título en documentos en español.

Se indica en estos casos, en la primera columna la identificación del archivo, en la segunda columna el Título que debería extraerse (a partir de la inspección manual de cada documento), en la tercera columna el Título extraído por el generador de metadatos, la cuarta columna muestra el grado de coincidencia entre ambos títulos, y las dos últimas columnas muestran

los valores de cobertura y de precisión según los lineamientos indicados en la Sección 2.1.1. El grado de coincidencia consiste en un valor numérico entre 0 y 2, con el siguiente criterio: valor 0 si el extractor no encontró nada o lo encontrado es basura; valor 1 si el extractor encontró el Título, pero extrajo texto adicional; y valor 2 si se obtuvo exactamente lo que se buscaba. Se aclara que en el caso de que el extractor no ha recuperado el título (Título Encontrado es vacío) no se ha transcripto el Título Real del documento. La Tabla 4 presenta algunos documentos de ejemplo en idioma español y la Tabla 5 los documentos de ejemplo en idioma inglés.

ID-Archivo	Título Real	Título Encontrado	Grado de coincidencia	Cobertura	Precisión
DI-1	CULTURAL SHIFTS Putting critical information literacy into practice	[ARTICLE] CULTURAL SHIFTS Putting critical information literacy into practice	1	100%	50%
DI-2	Marketing of the Library-Information Services	Marketing of the Library-Information Services	2	100%	100%
DI-3		---	0	0	N/A
DI-4		---	0	0	N/A
DI-5	Authorship, institutional and citation metrics ...	Osteoporos Int (2014) 25:1337–1343 DOI 10.1007/s00198-013-2603-3 ORIGINAL ARTICLE Authorship, institutional and citation metrics	1	100%	50%
DI-6	Libraries' Metadata as Data in the Era ... and PhD Dissertations for the Web of Data	Libraries' Metadata as Data in the Era ... and PhD Dissertations for the Web of Data *Manolis Peponakis*	1	100%	50%
DI-7	The AGROVOC Linked Dataset Editors	The AGROVOC Linked Dataset Editors*: Pascal Hitzler, ... anonymous reviewer*	1	100%	50%

Tabla 5. Ejemplo de asignación de valores para Título en documentos en inglés.

En la Tabla 5, en la columna Título encontrado, en las filas correspondientes a los documentos identificados como DI-1, DI-4, DI-5, DI-6 y DI-7, se utilizan puntos suspensivos para presentar el título en forma reducida, y se utiliza cursiva para indicar texto adicional al título extraído por el extractor. A modo de ejemplo, en el documento DI-6, además del título el extractor extrajo el texto *Manolis Peponakis*, motivo por el cual el grado de coincidencia se asigna con el valor 1 (el extractor extrajo el título y texto adicional), la cobertura es del 100% y la precisión es del 50%.

3.1.3 Metadato: Autores

Con los mismos subconjuntos de documentos tratados en las secciones anteriores, en las tablas siguientes (Tabla 6 y Tabla 7) se muestra la asignación de valores para los Autores de los documentos.

En la primera columna se muestra la identificación del archivo, en la segunda columna se muestra la cantidad de respuestas posibles, la tercera columna muestra la cantidad de respuestas encontradas que son válidas (según la inspección manual realizada), la columna cuatro la cantidad total de respuestas encontradas, y las dos últimas columnas muestran los valores de cobertura y de precisión según los lineamientos indicados en la Sección 2.1.1. La Tabla 6 presenta los documentos en idioma español y la Tabla 7 los documentos de ejemplo en idioma inglés.

ID-Archivo	Respuestas Posibles	Respuestas encontradas Válidas	Respuestas Totales encontradas	Precisión	Cobertura
DE-1	1	0	1	0	0
DE-2	1	0	0	N/A	0
DE-3	1	0	0	N/A	0
DE-4	1	1	1	100%	100%
DE-5	1	1	1	100%	100%
DE-6	1	0	0	N/A	0
DE-7	3	3	3	100%	100%

Tabla 6. Ejemplo de asignación de valores para Autores en documentos en español.

Archivo	Respuestas Posibles	Respuestas encontradas Válidas	Respuestas Totales encontradas	Precisión	Cobertura
DI-1	1	1	1	100%	100%
DI-2	7	5	5	100%	71,43%
DI-3	1	0	0	N/A	0
DI-4	1	0	0	N/A	0
DI-5	1	1	1	100%	100%
DI-6	2	2	2	100%	100%
DI-7	1	0	0	N/A	0

Tabla 7. Ejemplo de asignación de valores para Autores en documentos en inglés.

En la próxima subsección se presentan los resultados obtenidos sobre los corpus de 100 documentos en español y 100 documentos en inglés, para los tres metadatos analizados.

3.2 Resultados de la Experimentación

En las Tablas 8 y 9 se muestran para las Palabras Claves, la Precisión y la Cobertura respectivamente para documentos en idioma español y documentos en idioma inglés. Respecto a la Precisión se muestran los

porcentajes de documentos que alcanzaron porcentajes de precisión correspondientes al 100%, valores en los intervalos [75%, 100%), [50%, 75%), [25%, 50%) y [0%, 25%), distinguiéndolos de aquellos casos donde la Precisión no podía medirse (N/A) según los criterios indicados anteriormente. Respecto a la Cobertura se muestran, en la Tabla 9, los porcentajes de documentos que alcanzaron porcentajes de precisión correspondientes al 100%, valores en los intervalos [75%, 100%), [50%, 75%), [25%, 50%) y [0%, 25%).

Como se observa en ambas tablas, el porcentaje de documentos que alcanzaron un 100% de precisión en la recuperación de Palabras Clave en español (58%) es significativamente superior al alcanzado en el caso del idioma inglés (24%). Algo similar ocurre con la Cobertura, ya que la lograda en documentos en español (50%) duplica a la alcanzada en los documentos en inglés (24%). Cabe destacar, que en documentos en inglés y en español, se tiene que en el 84% y en el 81% respectivamente, se logra una Precisión superior al 50%, siendo muy buenos resultados, similares para ambos idiomas Respecto a la Cobertura, los resultados son mejores para el idioma inglés (80%) respecto al español (69%) cuando se evalúa una Cobertura superior al 50%.

% Precisión alcanzado	% Documentos que lo alcanzaron	
	Español	Inglés
100%	58%	24%
75% a 99,99%	7%	47%
50% a 74,99%	16%	13%
25% a 49,99%	4%	1%
< 25%	5%	3%
N/A	10%	12%

Tabla 8. Precisión para Palabras Claves.

% Cobertura alcanzado	% Documentos que lo alcanzaron	
	Español	Inglés
100%	50%	23%
75% a 99,99%	6%	47%
50% a 74,99%	13%	10%
25% a 49,99%	13%	7%
< 25%	18%	13%

Tabla 9. Cobertura para Palabras Claves.

Sobre el conjunto completo de los 100 documentos en español, como se observa en la Tabla 10, se obtuvo en promedio una Precisión de 75,81% y una Cobertura de 67,27%. Mientras que en el conjunto de documentos

en inglés, la Precisión fue del 73,04% y la Cobertura fue del 72,40%. Se puede concluir que aunque la Cobertura lograda en documentos en idioma inglés es levemente superior a la lograda en documentos en español (en un 5%), en este último idioma se logra una precisión algo superior (en un 3%).

Idioma del documento	Precisión	Cobertura
Español	75,81%	67,27%
Inglés	73,04%	72,40%

Tabla 10. Cobertura y Precisión para Palabras Claves según el idioma.

Respecto a los Títulos, los resultados obtenidos para Precisión y para Cobertura se muestran en las Tablas 11 y 12 respectivamente. El criterio seguido aquí es el ya descripto: N/A (No Aplicable) cuando el extractor no obtuvo resultados; si el resultado es parcial, por ejemplo si se extrajo el título y texto adicional, se asignó un valor 1 para la Cobertura y 0,5 para la Precisión; y si el resultado es exacto, se asignó un valor 1 para ambos indicadores.

% Precisión alcanzado	% Documentos que lo alcanzaron	
	Español	Inglés
100%	41%	31%
50%	8%	22%
N/A	51%	47%

Tabla 11. Precisión para Títulos.

% Cobertura alcanzado	% Documentos que lo alcanzaron	
	Español	Inglés
100%	49%	53%
0%	51%	47%

Tabla 12. Cobertura para Títulos.

Como se observa en las Tabla 11 y 12, aunque la cantidad de documentos en español (51%) donde no se extrajo el Título es superior a lo ocurrido en el idioma inglés (47%), se obtuvo un porcentaje del 41% para el idioma español para la extracción exacta del título, contra un 31% obtenido para el inglés.

Los resultados correspondientes a la evaluación de la Precisión y de la Cobertura para la extracción de Autores se presentan en las Tablas 13 y 14, respectivamente, donde se siguen los mismos lineamientos indicados para el caso de extracción de palabras claves.

% Precisión alcanzado	% Documentos que lo alcanzaron	
	Español	Inglés
100%	47%	50%
25% a 99,99%	0%	2%
< 25%	8%	1%
N/A	45%	47%

Tabla 13. Precisión para Autores.

% Cobertura alcanzado	% Documentos que lo alcanzaron	
	Español	Inglés
100%	44%	45%
25% a 99,99%	3%	7%
< 25%	53%	48%

Tabla 14. Cobertura para Autores.

Como se observa en la tablas anteriores, para el caso del idioma español, hay un 47% de archivos donde el extractor encontró total o parcialmente la respuesta esperada (C=1, P=1). En 8 archivos, aunque había una respuesta posible, esta no fue correcta (C=0 y P=0). Y en el 45% de los documentos el extractor no identificó al autor (esto es C=0 y P= N/A). Estos son los casos en que aunque había una respuesta posible, el extractor no la encontró. Para el caso del idioma inglés, los guarismos son similares.

Adicionalmente, se evaluó el reconocimiento, por parte del módulo Extractor, del Idioma del documento. En este metadato, se destaca que el idioma fue detectado correctamente en el 93% de documentos en español y en el 98% de los documentos en idioma inglés. Sólo en 7 documentos identificados como escritos en español, el idioma no fue detectado o fue detectado incorrectamente; y en idioma inglés esto ocurrió solo en el caso de 2 documentos.

Como conclusión general de la experimentación realizada, el módulo Extractor de Metadatos propuesto en este trabajo, se comporta en forma similar para el caso de los títulos y los autores, sea el documento en idioma inglés o en idioma español. Respecto al metadato Palabras clave, aunque la cobertura lograda en documentos en idioma inglés (72,40%) es levemente superior a la lograda en documentos en español (67,27%), en este último idioma se logra una precisión superior en casi un 3%. En general, la cobertura obtenida en los documentos en inglés es algo superior a la obtenida para los documentos en español, pero con una menor precisión.

Cabe resaltar que si para la evaluación de la Precisión se consideran solo los documentos en los que el extractor obtuvo metadatos (es decir, se analiza la precisión solo para los casos en donde el extractor produjo valores para el metadato en estudio), los valores alcanzados son promisorios. Para las Palabras claves, que es el metadato con mayores valores de cobertura (67,27% para el español y 72,40% para el inglés), para el idioma español se obtuvo una precisión del 85,70%, ligeramente superior a la precisión del 83% para los documentos en inglés. Para los títulos y para los autores, la cobertura se acerca al 50% tanto para los documentos en español como los documentos en inglés. Si se consideran solo aquellos documentos donde el extractor encontró respuestas para estos dos metadatos, la precisión para los títulos en español es del 96,08% y es del 100% para los documentos en inglés; y la precisión para los autores es del 85,45% para los documentos en español y del 96% para los documentos en idioma inglés.

4 Conclusiones

En este trabajo se presentó el desarrollo de un asistente para la carga de documentos en repositorios abordando la problemática del autoarchivo de objetos digitales educativos en un repositorio institucional. Para ello, se ha modificado el flujo de carga estándar de la plataforma DSpace, presentando un nuevo flujo para el depósito de objetos de modo que pueda integrarse un extractor de metadatos. Se propuso una arquitectura abierta de un módulo para la extracción automática de algunos metadatos de los documentos. Estos metadatos extraídos automáticamente son validados por el usuario en el proceso de descripción del objeto. Para diseñar el extractor se analizaron distintas herramientas de extracción y en particular se propuso utilizar la combinación de ParsCit+Alchemy, con la cual se logró incrementar la calidad de las palabras claves retornadas pasando de un 56% de resultados correctos obtenidos con Alchemy a un 70%.

Cabe destacar que el módulo extractor se ha diseñado de modo de poder integrar otras herramientas extractoras, siempre teniendo presente que una condición clave para este módulo es que produzca los resultados en tiempo real. Se ha desarrollado un prototipo en JAVA de este asistente y en este trabajo se presentan los resultados obtenidos sobre dos corpus de documentos, uno de documentos en español y otro de documentos en inglés obtenidos del repositorio e-LIS (eprints.rclis.org). Se evaluó el reconocimiento, por parte del módulo Extractor, del Idioma del documento, siendo detectado correctamente en el 93% de documentos en español y en el 98% de los documentos en inglés. Para el caso de las Palabras claves, la precisión obtenida para los documentos en español es de 75,81% y es de 73,04% para los documentos en inglés. Para el metadato

Título, para ambos idiomas la cobertura se acerca al 50%, y se obtuvo un porcentaje del 41% para el idioma español para la extracción exacta del título, contra un 31% obtenido para el inglés. Para los Autores, tanto la cobertura como la precisión son ligeramente inferiores al 50% en ambos idiomas. Cabe resaltar que si para la evaluación de la precisión se consideran solo los documentos en los que el extractor obtuvo metadatos, los valores alcanzados para esta métrica resultan promisorios: para las Palabras claves, que es el metadato con mayores valores de cobertura se obtiene una precisión del 85,70% para los documentos en español, ligeramente superior al 83% para los documentos en inglés; para los títulos, la precisión para documentos en español es del 96,08% y es del 100% para los documentos en inglés; y para los autores es del 85,45% para los documentos en español y del 96% para los documentos en idioma inglés.

Como conclusión de esta experimentación puede evaluarse que los resultados preliminares son promisorios y que el módulo extractor propuesto, cuando recupera información de los metadatos buscados lo hace con muy buena precisión. Como trabajo futuro, se plantea proponer y experimentar con otros posprocesamientos de extracción que mejoren la cobertura en los documentos donde no se pudo extraer información para algunos de los metadatos considerados. Este problema se debe principalmente, a los distintos formatos con que están editados los documentos.

Este módulo extractor, se va a implementar en el Repositorio RepHip de la Universidad Nacional de Rosario donde se podrá realizar una experimentación más detallada. De esta forma se espera ayudar al usuario en el proceso de carga de objetos digitales educativos, disminuyendo así su trabajo y mejorando la cantidad y la calidad de los metadatos cargados.

Agradecimientos

Este trabajo ha sido parcialmente financiado por la Red CYTED RIURE: Red Iberoamericana para la Usabilidad de Repositorios Educativos, por el proyecto de Redes VII "Red para la Integración de Universidades en el uso de TIC para la Inclusión en la Educación Superior" y por el Proyecto LATIn: Latin American open Textbook Initiative, Alfa III DCI-ALA/19.09. 01/11/21526/279-155/ALFA III (2011)-52.

Referencias

[1] D. Wiley. Connecting Learning Objects to Instructional Design Theory: A definition, a metaphor, and a taxonomy, en D. A. Wiley

(ed.) Instructional Use of Learning Objects. Editorial Association for Instructional Technology, 2002.

[2] M. Sonntag. Metadata in E-Learning Applications: Automatic Extraction and Reuse, in Christian Hofer, Gerhard Chroust (Eds.): IDIMT-2004. 12th Interdisciplinary Information Management Talks, pp. 219-231, Universitätsverlag Rudolf Trauner, Linz, Austria, 2004.

[3] V. Gerling. Un Sistema Inteligente para Asistir la Búsqueda Personalizada de Objetos de Aprendizaje, Degree Thesis on Computer Science, National University of Rosario, Argentina, 2009. Disponible en www.fceia.unr.edu.ar/lcc/t523/tesina.php?campo1=21

[4] A. Casali, C. Deco, C. Bender, S. Fontanarrosa y C. Sabater. "Asistente para el Depósito de Objetos en Repositorios con Extracción Automática de Metadatos". En XV Simposio Internacional de Tecnologías de la Información y las Comunicaciones en la Educación (SINTICE 2013), pp 133-136. Madrid, España, setiembre 2013.

[5] S. Beltramone. Sistema de recopilación de objetos digitales en dominios restringidos. Tesina de Grado, Licenciatura en Ciencias de la Computación, Universidad Nacional de Rosario, 2014.

[6] A.Casali, C.Deco and S.Beltramone. "Automatic Gathering of Educational Digital Resources to Populate Repositories", in Proc. Of Interacción 2014, ACM Digital Library, Article No. 85. PublisherACM New York, NY, USA. Puerto de la Cruz, Tenerife, España, setiembre 2014.

[7] P. San Martín, P. Bongiovani, A. Casali, C. Deco. Socio-technological perspectives for Open Access Repositories development in the context of public universities in the central-eastern Argentina. PKP Scholarly Publishing Conference, DF Mexico. 2013.

[8] A. Esuli and F. Sebastiani, Evaluating Information Extraction. M. Agosti et al. (Eds.): CLEF 2010, LNCS 6360, pp. 100–111, Springer-Verlag Berlin Heidelberg 2010.

[9] R. Grishman, Information Extraction: Capabilities and Challenges. Notes prepared for the 2012 International Winter School in Language and Speech Technologies Rovira i Virgili University. Tarragona, Spain. 2012

[10] R. Baeza-Yates, B. Ribeiro-Neto (eds.). Modern Information Retrieval. New York. ACM Press, 1999.

[11] F. Martínez Méndez. Recuperación de información: modelos, sistemas y evaluación. Murcia: KIOSKO JMC, 2004.

[12] T. Pire, B. Espinase, A. Casali, and C. Deco. "Automatic Extraction of Learning Objects Metadata for Recommendation: A Comparative Study", in Proc. XIV Congreso Internacional de Informática en la Educación (InforEdu 2011). La Habana, Cuba, febrero 2011.

[13] I. Witten, G. Paynter, E. Frank, C. Gutwin, and C. Nevill-Manning. "KEA: Practical automatic keyphrase extraction", in Proc. of the fourth ACM conference on Digital libraries. pp 254-255. ACM New York, NY, USA 1999.

[14] J. Beel, B. Gipp, S. Langer, M. Genzmehr, E. Wilde, A. Nürnberger, and J. Pitman. "Introducing Mr. DLib, a Machine-readable Digital Library", in Proc. 11th ACM/IEEE Joint Conference on Digital Libraries (JCDL'11), ACM. 2011.

[15] I. Councill, C. Giles, and M. Kan. "ParsCit: An open-source CRF reference string parsing package", in International Language Resources and Evaluation. Ed. European Language Resources Association. 2008.