UNIVERSIDAD TECNOLOGICA NACIONAL

FACULTAD REGIONAL SANTA FE

MAESTRÍA EN INGENIERÍA EN SISTEMAS DE INFORMACIÓN

Tesis de maestría

AMELOIR: ALGORITMO PARA LA EXTRACCIÓN AUTOMÁTICA DE METADATOS A PARTIR DE OBJETOS DE APRENDIZAJE EN UN REPOSITORIO INSTITUCIONAL

Ing. Adriana Constanza Pinilla Gómez

Directora: Dra. María de los Milagros Gutiérrez

Co-directora: Dra. Luciana Ballejos

Santa Fe Agosto 2017

Se presenta esta tesis en cumplimiento de los requisitos exigidos por la Universidad Tecnológica Nacional para la obtención del grado académico de magister en Ingeniería en sistemas de Información

AMELOIR: ALGORITMO PARA LA EXTRACCIÓN AUTOMÁTICA DE METADATOS A PARTIR DE OBJETOS DE APRENDIZAJE EN UN REPOSITORIO INSTITUCIONAL

por

Ing. Adriana Constanza Pinilla Gómez

Directora: Dra. María de los Milagros Gutiérrez

Co-directora: Dra. Luciana Cristina Ballejos

Jurado de Tesis:

Dra. Lucila Romero

Dr. Jorge Roa

Dr. Mariano Rubiolo

Santa Fe Agosto 2017

Contenido

Capítulo 1

Introducción

En el dominio de la educación, gran cantidad y diversidad del material utilizado en el proceso de enseñanza-aprendizaje se encuentra publicado, pero disperso en Internet. La recuperación de dicho material se realiza haciendo uso de buscadores, tales como Google (Google, 1998), pero actualmente existen otras formas más estructuradas de acceder a la información. Bajo este contexto aparecen cuatro conceptos fundamentales que enmarcan esta realidad educativa: *objetos de aprendizaje*, *metadatos*, *estándares* y *repositorios institucionales*.

Los *objetos de aprendizaje* (de ahora en adelante denominados OA), son "cualquier recurso digital que puede ser reutilizado para la enseñanza" (Wiley, 2001). Estos objetos pueden adquirir formas muy diversas y presentarse en diferentes formatos. Además, para poder contar con la posibilidad de ser reutilizados, actualizados, combinados, separados, referenciados y sistematizados, deben tener ciertas características, entre las cuales, las más significativas son: *accesibilidad*, *reusabilidad* e *interoperabilidad* (Polsani, 2006).

Accesibilidad se refiere a que el OA pueda ser accedido por diferentes tipos de usuarios (estudiantes, docentes, investigadores). Por *reusabilidad* se entiende que el OA pueda ser usado una y otra vez dentro de diferentes contextos educativos y/o para diversas finalidades. Finalmente, *interoperabilidad* es la capacidad del OA para ser utilizado en diferentes repositorios, herramientas o plataformas, independientemente de sus características técnicas. Otras características a tener en cuenta son: durabilidad, independencia, escalabilidad y adaptabilidad, flexibilidad, versatilidad y funcionalidad, interactividad y granularidad, así como la capacidad para ser evaluado, manejado, recuperado e intercambiado.

La mayoría de estas características están relacionadas con otros dos conceptos de gran importancia para el almacenamiento, distribución y reutilización de los OA: *metadatos* y *estándares*.

Los *metadatos* son un conjunto de atributos o etiquetas que describen las principales características de un OA y proporcionan información adicional sobre el mismo. La información almacenada en los metadatos es fundamental para garantizar el éxito en la interconexión entre repositorios y facilitar el desarrollo de sistemas de búsqueda, tales como los sistemas recomendadores. Dicha información es un aspecto clave en la calidad y precisión de los objetos retornados por estos sistemas, para satisfacer las necesidades de búsqueda de los usuarios y favorecer la reutilización de los OA.

La calidad y pertinencia de los metadatos definidos para los OA, se evalúa a la luz del cumplimiento de *estándares de metadatos* tales como *DublinCore* (DCMI, Dublin Core Metadata Initiative, 1995) e *IEEE LOM* (IEEE, 2002), que utilizan distintas categorías para describir el contenido del objeto (título, autor, palabras claves, idioma, etc.), e incluso, como en el caso de LOM, permiten describir aspectos educacionales de los mismos (nivel educativo, complejidad, etc.).

El concepto Repositorios Institucionales de Objetos de Aprendizaje (de ahora en adelante denominados RI) reúne las nociones de OA, metadatos y estándares, entendiéndose al mismo como una gran colección de OA, estructurada como una base de datos, con metadatos asociados generalmente bajo el cumplimiento de algún estándar y que, en la mayoría de los casos, se puede encontrar en la Web (Casali, Gerling, Deco, & Bender, 2009).

Gracias a la creación de estos RI que reúnen o almacenan en un solo lugar y de manera ordenada diversos OA, se hace posible que gran cantidad de usuarios tengan acceso a la información que allí se encuentra. Esto permite resolver, en alguna medida, la problemática donde muchas veces el material educativo disponible no se aproveche de la mejor forma y además, que el conocimiento no sea difundido, promovido ni reutilizado.

Uno de los grandes retos en el área es propiciar el uso de los RI a través de la búsqueda y consulta de material educativo. Para esto, es importante contar con una buena descripción de los OA que conforman el repositorio, con el fin de promover el éxito en las búsquedas a partir de, por ejemplo, búsquedas inteligentes o recomendadas, que hacen referencia a aquellos sistemas de búsqueda que son capaces de seleccionar, de forma automática y personalizada, el material que mejor se adapte al perfil, preferencias o necesidades del usuario (Casali, Gerling, Deco, & Bender, 2011). Un aspecto clave en la recuperación y pertinencia de los OA

retornados es la calidad de los metadatos descriptivos, que proporcionan información fundamental para la reutilización de los mismos.

1.1 Fundamentación

Uniendo los conceptos de *repositorio institucional* (RI), *objeto de aprendizaje* (OA), *metadatos* y *estándares*, se establece un marco de referencia donde es posible determinar la importancia del *acceso abierto* y masivo a recursos digitales educativos.

El concepto de *acceso abierto* incluye cualquier iniciativa, proyecto o acción que favorezca y promueva el acceso libre a través de Internet a las publicaciones científicas (Melero, 2007). Según Budapest Open Access Initiative, el acceso abierto es la disponibilidad gratuita de la literatura en Internet que permite que cualquier usuario pueda leer, descargar, copiar, imprimir y/o distribuir la información sin ninguna barrera financiera, legal o técnica, siendo la única restricción sobre la distribución y reproducción la de dar a los autores control sobre la integridad de su trabajo y el derecho a ser citado y reconocido adecuadamente (Budapest Open Access Initiative, 2002).

Los orígenes fundacionales del acceso abierto tienen poco más de una década, aunque las primeras experiencias se encuentran más de veinte años atrás. Fue a principios de 1990 que surgieron las primeras iniciativas para crear archivos o repositorios abiertos de documentos especializados, con el fin de facilitar el acceso a los contenidos, hasta ese momento sólo disponibles para los que pudiesen pagar. Desde entonces, el movimiento ha crecido y evolucionado a nivel mundial, y son cada vez más las instituciones académicas que apoyan la creación de repositorios o iniciativas de este tipo.

Existen dos estrategias para hacer posible el acceso abierto (De Volder, 2012):

1. *Vía dorada*: publicar en revistas de acceso abierto.
2. *Vía verde*: autoarchivar trabajos de investigación en repositorios temáticos o institucionales.

Más aún el acceso abierto, que surge a raíz de la digitalización y del incremento en los costos de suscripción a revistas y de compra libros en general, tiene como fin aumentar la visibilidad, uso, difusión e impacto de la investigación. Así mismo, favorecer la educación y el desarrollo,

romper las barreras de conocimiento entre países ricos y en vías de desarrollo, y recuperar parte de la inversión dedicada a la investigación científica.

La consolidación de la Red Federada de Repositorios Institucionales de Publicaciones Científicas en América Latina busca dar un valor agregado a la tendencia mundial de acceso abierto al material educativo y, por lo tanto, al conocimiento existente en diversas áreas. En este sentido, esta red ha venido trabajando para lograr acuerdos y establecer políticas regionales con el fin de promover la visibilidad de los RI y contribuir con la difusión del conocimiento, principalmente en América Latina.

Los RI se reconocen como fuente importante de información científica y académica, y constituyen la base fundamental hacia la democratización del conocimiento. Además, hacen posible la visibilidad y difusión mundial de los recursos publicados, aumentan la posibilidad de que los OA sean conocidos y citados por docentes, estudiantes e investigadores, y garantizan, de cierta manera, la perdurabilidad, reconocimiento y contexto académico de la información digital.

Por otra parte, el autoarchivo -proceso de almacenamiento del OA en un repositorio por el mismo autor o cualquier usuario, con el fin de que estén disponibles en acceso abierto a través de Internet- trae consigo desafíos, debido a la falta de conocimiento del significado e importancia de los metadatos y su respectiva correspondencia con la información del OA. De aquí surge la importancia de poder garantizar que los metadatos asociados a un OA lo describan de forma adecuada y verídica, de tal manera que sea posible identificar claramente al objeto, lograr una buena interconexión entre repositorios y facilitar el desarrollo de sistemas de búsqueda; una forma de lograrlo es a través de extractores automáticos o semiautomáticos, como herramientas que se encargan de obtener un conjunto de metadatos a partir del contenido de un OA, mediante el uso de distintas técnicas y recursos.

Aunque existen diferentes propuestas para la extracción automática de metadatos a partir de OAs, cada una cuenta con sus propios objetivos y arquitectura. Sin embargo, hasta el día de hoy, esta área no ha sido lo suficientemente considerada, a pesar de la gran importancia e interés que reviste a nivel de RI. Esto es así, principalmente, porque favorece la precisión de las búsquedas y permite la recuperación de aquellos objetos que mejor satisfagan las necesidades de información del usuario, teniendo en cuenta sus características y preferencias individuales.

Desde el punto de vista tecnológico, los metadatos constituyen una parte fundamental no sólo de los objetos de aprendizaje para que éstos puedan ser encontrados y reutilizados, sino de los repositorios en sí, ya que hace posible que aumenten los niveles de confianza en la utilización de estas herramientas de búsqueda y consulta. Por otro lado, a nivel educativo, los metadatos hacen posible que se pueda evaluar la pertinencia y calidad de los objetos de aprendizaje incluidos en los resultados de la búsqueda, a partir de un fin educativo concreto y de acuerdo al perfil del usuario que realiza la búsqueda.

Si bien hay algunos estudios en lo que respecta a extracción automática de metadatos en OA, es largo el camino que queda por recorrer, ya que actualmente no existe una propuesta que contemple y afronte los siguientes desafíos:

- **Aplicación de estrategias de inteligencia artificial**

Indudablemente, para la creación de nuevos algoritmos para la extracción automática de un conjunto de metadatos apropiados a partir de OA en un RI de acceso abierto, se requiere la aplicación de estrategias de inteligencia artificial tales como reglas de mapeo directo, reglas heurísticas de mapeo, herramientas de procesamiento de lenguaje natural, ontologías, etc., lo que se convierte en un verdadero desafío de investigación y aplicación de la ingeniería.

- **Automatización de la extracción de metadatos**

La generación semiautomática de metadatos, donde hay validación/corrección de la información asociada al OA por parte del usuario que está realizando el proceso de autoarchivo en el repositorio, aumenta el grado de imprecisión, datos incompletos, inconsistencias y discrepancias de la información almacenada en el repositorio. Automatizar esta actividad se convierte en un desafío para hacer posible que dichos objetos queden mejor descriptos y de esta manera se garantice la calidad de la búsqueda, recuperación y reutilización de los mismos.

- **Involucrar la mayor cantidad y variedad de categorías de OA**

Es necesario que los nuevos algoritmos contemplen tanto aquellos tipos de archivos conocidos y de uso común, como documentos de texto y PDF, y otros tales como los archivos tipo música e imágenes que tienen un mayor grado de dificultad. Aún más, el soporte debe brindarse no sólo en el momento de ser autoarchivados y clasificados sino también en el de ser incluidos en los resultados de las búsquedas, precisamente porque los metadatos asociados a los mismos no son claramente identificados.

- **Integración de estándares de metadatos**

Una de las grandes ventajas es el avance que se tiene en cuanto a la definición de estándares para metadatos, que facilitan de cierta manera tanto el almacenamiento como la recuperación de OA en los repositorios; sin embargo, hasta el momento ninguna de las propuestas de sistemas de extracción semi-automática de metadatos hace uso integrado de los estándares Dublin Core (DCMI, Dublin Core Metadata Initiative, 1995) e IEEE LOM (IEEE, 2002), por lo tanto, tener en cuenta esta perspectiva de trabajo en la generación de nuevos algoritmos y herramientas para la extracción automática de metadatos es un interesante desafío, como una forma de aprovechar los grandes esfuerzos que se han hecho hasta el momento para brindar la posibilidad de estandarizar y unificar la información que se extrae y almacena de los OA.

- **Interoperabilidad entre los diferentes RI de acceso abierto**

El uso de estándares para metadatos en la definición de nuevos algoritmos para la extracción automática de metadatos en OA, también hará posible la compatibilidad e interoperabilidad entre diferentes RI, para así ampliar el rango de búsqueda y la reutilización de los OA y contribuir significativamente con el acceso abierto al conocimiento científico-educativo que se encuentra disponible en las diferentes instituciones educativas y en Internet y que puede llegar a ser muy valioso y útil dentro del proceso de enseñanza-aprendizaje.

Por lo tanto, el desarrollo de nuevos algoritmos y sistemas de extracción automática parece ser un paso muy importante para afrontar estos desafíos, garantizando que se extraiga del OA la mayor cantidad posible de información de alta calidad. Esto permitirá que el OA quede mejor descripto, evitando dejar a criterio del usuario la inclusión de información valiosa para la recuperación y reutilización de estos recursos digitales.

1.2 Objetivos

1.2.1 General

El objetivo principal de la presente tesis es el diseño de un nuevo algoritmo que permita la extracción automática de metadatos, a partir de objetos de aprendizaje en un RI de acceso abierto, considerando un conjunto de metadatos preestablecidos.

1.2.2 Específicos

Para lograr el objetivo general definido previamente será necesario lograr los siguientes objetivos particulares de manera progresiva:

1. Realizar un diagnóstico sobre el manejo de los metadatos y recuperación de objetos de aprendizaje en RI.
2. Analizar y comparar las diferentes estrategias de extractores de metadatos existentes identificando el formato de archivos sobre los que se aplican.
3. Identificar las categorías de OA que existen en los diferentes repositorios y determinar para cada categoría los metadatos que serán extraídos.
4. Definir los formatos de archivos con los cuales se trabajará para el diseño del algoritmo de extracción de metadatos a desarrollar, en base a aquellos que se almacenan con mayor frecuencia en los RI de Acceso Abierto.
5. Proponer técnicas de extracción a utilizar para la obtención de metadatos dependiendo de cada categoría de OA, de los metadatos identificados para extraer y de los formatos de archivos que serán tenidos en cuenta para el diseño del algoritmo.

1.3 Aportes de la tesis

El principal aporte de esta tesis es el diseño e implementación de AMELOIR (Automatic Metadata Extracción Learning Object Institutional Repository), un nuevo algoritmo para la extracción automática de metadatos en RI utilizando técnicas de procesamiento de lenguaje natural e inteligencia artificial. AMELOIR fue incorporado en la plataforma DSpace alterando el proceso de almacenamiento (ver Capítulo 4), de tal manera que al ser cargado un archivo para almacenar, se invoca al extractor para que obtenga automáticamente los metadatos. Éstos se presentan al usuario en la etapa de verificación de metadatos, para que sean validados y completados en caso de que sea necesario.

Con el desarrollo de nuevos algoritmos y sistemas de extracción automática de metadatos como el propuesto en esta tesis, se reducen en gran medida los inconvenientes que presenta el autoarchivo con respecto al uso de los metadatos que no contienen información, o bien, cuando la información que poseen es incorrecta o de baja calidad.

Dichos inconvenientes se presentan en gran parte por desconocimiento de valor y significado de los metadatos por parte de quien archiva el OA en el repositorio, por no contar con una herramienta adecuada de extracción automática, por la falta de selección por parte de los desarrolladores de repositorios de un conjunto adecuado de metadatos (obligatorios y

opcionales) para describir los OA, así como por la diversidad de normas y el soporte de motores de búsqueda.

Además, en el algoritmo que se va a proponer se afrontan los siguientes desafíos con respecto a la extracción automática de metadatos de OA en RI de acceso abierto:

- Se aplican herramientas de procesamiento de lenguaje natural y expresiones regulares para la extracción de metadatos.
- Se reduce, hasta donde sea posible, la intervención del usuario en el proceso de extracción de metadatos, para disminuir el grado de imprecisión, incompletitud, inconsistencias y discrepancias de la información almacenada en el repositorio.
- Se contempla el procesamiento de archivos en formato Word y PDF.
- Se hace uso integrado de estándares de metadatos Dublin Core e IEE LOM, como una forma de aprovechar los grandes esfuerzos que se han hecho hasta el momento para brindar la posibilidad de estandarizar y unificar la información que se extrae y almacena de los OA, haciendo posible la compatibilidad e interoperabilidad entre diferentes RI.

Al ser AMELOIR un algoritmo de extracción de metadatos aplicable a cualquier RI implementado en DSpace, el software más utilizado a nivel global en lo que a RI de universidades y centros de investigación respecta, es posible adaptar dicho algoritmo para ser utilizado en cualquier otro repositorio que cumpla con estas características, haciendo factible su inserción en mercados externos. El atractivo del algoritmo está en permitir el enriquecimiento de las descripciones de los recursos educativos y de esta manera, optimizar las funcionalidades de búsqueda implementadas en cada uno de los repositorios. Adicionalmente, se estarían sustituyendo herramientas que trabajen de manera independiente al RI o con un menor nivel de detalle, y que a su vez podrían ser privativas o generar algún costo de licenciamiento.

A partir del desarrollo y adaptación de AMELOIR, se hace factible sustituir la importación de material educativo pago, tanto físico como virtual, debido a que se incentiva la reutilización del material producido en la misma institución y la colaboración entre, por ejemplo, universidades y grupos de investigación de distintos países. De esta forma, se enriquecen las funcionalidades disponibles del RI utilizado, logrando contar con un producto de gran valor tecnológico, a la vez que se reutiliza y agrega valor promoviendo el conocimiento generado internamente.

Por otra parte, con la implementación de este algoritmo se contribuye de manera significativa al desarrollo del RI del CIDISI (Centro de Investigación y Desarrollo de Ingeniería en Sistemas de Información, Universidad Tecnológica Nacional, Facultad Regional Santa Fe), el cual está en su etapa inicial de implementación, y a través de éste último se da impulso a proyectos futuros a partir de la distribución, visibilidad y reutilización del conocimiento logrado en investigaciones previas.

Como resultado de las investigaciones, se presentó el siguiente trabajo en congreso:

Pinilla A.; Gutiérrez, M.; Ballejos L. Extracción Automática de Metadatos a partir de Objetos de Aprendizaje en un Repositorio Institucional: Estado del Arte. Anales 43 JAIIO, Simposio Argentino de Tecnología y sociedad. Buenos Aires 2014 p 67 - 82 -. issn 2362-5139. (2014).

1.4 Organización de la tesis

Esta tesis se encuentra organizada de la siguiente manera: en el capítulo 2 se presenta el marco teórico, donde se definen los conceptos de RI, acceso abierto, metadatos y estándares, y se realiza una descripción de las técnicas procesamiento de lenguaje natural (NLP) utilizadas en el algoritmo de extracción propuesto.

El capítulo 3 analiza el estado del arte de extractores de metadatos, las técnicas usadas y aplicabilidad de los mismos, se los analiza a la luz de tres aspectos importantes que se deben tener en cuenta en el momento de elegir o diseñar un sistema de este tipo: los tipos de archivos a procesar, los metadatos extraídos y las técnicas y recursos utilizados para realizar la extracción.

En el capítulo 4 se presenta AMELOIR, el algoritmo de extracción propuesto, dando una explicación detallada de cada uno de los pasos que se tuvieron en cuenta para construirlo. En este mismo capítulo se presenta casos de estudio donde se aplica el algoritmo propuesto en algunos archivos y se presentan los resultados obtenidos.

En el capítulo 5 se presentan las conclusiones y trabajos futuros.

Por último se encuentran la bibliografía y anexos.

Capítulo 2

Marco Teórico

2.1 Repositorios Institucionales y Acceso Abierto

La creación de Repositorios Institucionales de Acceso Abierto constituye un fundamento importante para la preservación, promoción y difusión del conocimiento, haciendo posible que muchos usuarios, entre los que se encuentran estudiantes, docentes e investigadores, puedan acceder a gran cantidad y diversidad de material educativo que puede ser reutilizado en el proceso de enseñanza-aprendizaje.

A continuación se realiza un diagnóstico sobre la situación actual de los RI de acceso abierto tanto en Argentina como en Colombia, describiendo para cada caso, las principales características de aquellos que se consideran más representativos.

2.1.1 Repositorios Institucionales en Argentina

En Argentina la iniciativa de crear una red de repositorios digitales comenzó a materializarse a mediados de 2009 en el Ministerio de Ciencia, Tecnología e Innovación Productiva (MinCyT); posteriormente, mediante Resolución Ministerial N°469/11 del 17 de Mayo de 2011, se creó el Sistema Nacional de Repositorios Digitales (SNRD), conjuntamente con el Consejo Interinstitucional de Ciencia y Tecnología (CICyT), a través de sus representantes en el Consejo Asesor de la Biblioteca Electrónica de Ciencia y Tecnología. El SNRD tiene como propósito conformar una red interoperable de repositorios digitales en ciencia y tecnología, a partir del establecimiento de políticas, estándares y protocolos comunes a todos los integrantes del Sistema (Ministerio de Ciencia, 2002).

Así mismo, el 23 de Mayo de 2012, la Cámara de Diputados de la Nación Argentina dio media sanción al proyecto de Ley que obliga a las instituciones del Sistema Nacional de Ciencia y Tecnología que reciban financiamiento del Estado Nacional, a crear repositorios digitales institucionales de acceso abierto y gratuito en los que se depositará la producción científico-

tecnológica nacional (Honorable Cámara de Diputados de la Nación, 2010). Finalmente, el 13 de noviembre de 2013 la cámara de senadores la aprueba convirtiéndola en ley.

La producción científica que es publicada en los repositorios digitales abarca trabajos técnico-científicos, tesis académicas, artículos de revistas, entre otros, que sean resultado de la realización de actividades de investigación financiadas con fondos públicos, ya sea, a través de sus investigadores, tecnólogos, docentes, becarios postdoctorales y estudiantes de maestría y doctorado. La Ley establece, además, la obligatoriedad de publicar los datos de investigación primarios luego de 5 años de su recolección, para que puedan ser utilizados por otros investigadores.

El objetivo es que la producción científica financiada por la sociedad sea accesible a quien lo solicite. Por supuesto, aquellas investigaciones que requieran confidencialidad no deben ser publicadas. Por otra parte, Alejandro Ceccatto -secretario de Articulación Científico Tecnológica del Ministerio- destaca que la propiedad intelectual y las patentes están protegidas y no se ven afectadas por esta forma de democratización de la información científica (Honorable Cámara de Diputados de la Nación, 2010).

La interoperabilidad de los repositorios digitales que deberán crear las instituciones será diseñada por el SNRD a fin de garantizar el acceso libre, gratuito y universal desde un único portal. Según los fundamentos del proyecto, el modelo de acceso abierto a la producción científico tecnológica implica que los usuarios de este tipo de material pueden, en forma gratuita, leer, descargar, copiar, distribuir, imprimir, buscar o enlazar los textos completos de los artículos científicos, y usarlos con propósitos legítimos ligados a la investigación científica, a la educación o a la gestión de políticas públicas, sin otras barreras económicas, legales o técnicas que las que suponga Internet en sí misma.

A nivel global, cabe destacar que en la región solo Perú posee una ley de Acceso Abierto sancionada en el año 2013, que convirtió al país en el segundo de América Latina, después de Argentina, en elevar una legislación nacional al respecto. En el caso de Estados Unidos, la obligatoriedad de publicar las investigaciones sólo alcanza a aquellas financiadas con fondos públicos a través de sus Institutos Nacionales de Salud (NIH). Finalmente, la Comisión Europea también promueve el acceso abierto pero todavía con iniciativas aisladas.

2.1.2 Repositorios Institucionales en Colombia

En lo que respecta a lo existente en el área en Colombia (Campo Saavedra, Martínez Barrios, Ruíz Rodgers, & Rendón Osorio, 2012), y teniendo en cuenta que es evidente que las Tecnologías de la Información y las Comunicaciones (TICs) juegan y jugarán un rol protagónico en el fortalecimiento de la capacidad de los sistemas educativos y en el mejoramiento de su calidad, es constante el impulso que desde el Ministerio de Educación Nacional se da para mejorar las condiciones y los servicios de la infraestructura tecnológica nacional y promover su apropiación y uso por parte de las comunidades educativas. Inicialmente, este apoyo se dio desde el Programa Nacional de Uso de Medios y TIC (2003 – 2011) y, actualmente, a través de la consolidación del Sistema Nacional de Innovación Educativa con Uso de TICs, que lidera la Oficina de Innovación Educativa con Uso de Nuevas Tecnologías.

A través de este sistema, y con el apoyo del Departamento Administrativo de Ciencia, Tecnología e Innovación (COLCIENCIAS), el Ministerio de Educación Nacional y el proyecto Biblioteca Digital Colombiana (BDCOL) articula los repositorios institucionales de las universidades colombianas. Específicamente, en el año 2011 comenzó a impulsar el diseño e implementación de la *Estrategia Nacional de Recursos Educativos Digitales Abiertos* dirigida a Educación Superior, que busca contribuir a mejorar las condiciones de acceso a la información y al conocimiento por parte de las comunidades educativas, a fortalecer la capacidad del uso educativo de las TICs, a fomentar una cultura en torno a la colaboración y cooperación para promover el intercambio, reutilización, adaptación, combinación y redistribución de recursos educativos, y a consolidar una amplia oferta nacional de recursos de acceso público, que aporte al mejoramiento de la calidad en la educación, además de articularse con los planteamientos recogidos por la UNESCO en la Declaración de París de Junio de 2012.

A su vez, esta Estrategia Nacional responde a los documentos Visión 2019, el Plan Decenal de Educación 2006-2016, a los propósitos generales de la Educación de Calidad como camino a la prosperidad, y a los mejores intereses de la nación para fortalecer el acceso a la educación y el conocimiento. La Estrategia Nacional pretende involucrar a todas las Instituciones de Educación Superior del país para que aporten Recursos Educativos Digitales Abiertos, de la más alta calidad, bien sea para Educación Superior o para el cumplimiento de metas sociales y culturales de las mismas.

Tabla 2.1. Historia del acceso abierto y repositorios institucionales en Colombia.

Año	Acontecimiento	Descripción
2005	Creación del Portal Educativo Colombia Aprende	Creado con el fin de converger a la comunidad académica en torno a una oferta de contenidos, herramientas y servicios de oportuna respuesta a los usuarios. Cuenta con servicios transversales que lo complementan, como el acceso a contenidos de proveedores externos, recursos digitales, bases de datos y estadísticas nacionales y mundiales. Para el año 2005 contaba con 4797 recursos de origen nacional y 2814 recursos de fuentes internacionales (Ver Anexo 2).
2006	Proyecto de Catalogación de Objetos de Aprendizaje	Este proyecto tuvo como origen el marco del Foro de Investigadores de la *Red Iberoamericana de Informática Educativa* (RIBIE), celebrado en Santa Marta, donde se compartieron opiniones, se generaron discusiones y se desarrollaron mesas de trabajo. La iniciativa de *Objetos de Aprendizaje* se dividió en dos etapas distintas pero complementarias: la primera, de carácter conceptual y la segunda de naturaleza práctica, orientadas a la generación de una red federada de repositorios institucionales vinculados al *Banco Nacional de Objetos de Aprendizaje*.
2007	Conformación de la Red de Catalogación de Objetos de Aprendizaje en Colombia Aprende	Se conformó con doce (12) instituciones, más el banco del Ministerio de Educación Nacional MEN, distribuidas en todo el territorio nacional.
	Creación de la Red Nacional Académica de Tecnología Avanzada (RENATA) [26]	RENATA es la red de tecnología avanzada que conecta, comunica y propicia la colaboración entre las instituciones académicas y científicas de Colombia con las redes académicas internacionales y los centros de investigación más desarrollados del mundo. Está integrada por tres miembros del Gobierno (Ministerio de Educación Nacional, Ministerio de Tecnologías de la Información y las Comunicaciones y Departamento Administrativo de Ciencia, Tecnología e Innovación COLCIENCIAS) y ocho Redes Académicas Regionales.
	Creación de la Biblioteca Digital Colombiana BDCOL	Permite agrupar y dar visibilidad nacional e internacional a la producción académica, científica, cultural y social de acceso abierto de instituciones educativas, centros de investigación, centros de documentación, organismos gubernamentales y no gubernamentales, archivos y bibliotecas en general que estén comprometidas con el progreso educativo del país (Ver Anexo 2). BDCOL (mediante RENATA) está articulado al Proyecto BID – Bienes Públicos Regionales, como parte de la "Estrategia Regional, Marco de Interoperabilidad y Gestión para una Red Federada Latinoamericana de Repositorios Institucionales de Documentación Científica" que se da entre los siguientes países: Colombia, México, Argentina, Chile, Venezuela, Perú, Ecuador.
2010	Catalogación de Objetos de Aprendizaje en Instituciones de Educación Superior	La participación de grupos institucionales en el proceso de Catalogación de Objetos de Aprendizaje, estuvo integrada por profesionales cuyos perfiles poseen fortalezas tecnológicas y pedagógicas principalmente.
	Elaboración del Contrato de Ciencia y Tecnología Nº 441	Este contrato se suscribió entre el Ministerio de Educación Nacional y la Corporación Red Nacional Académica de Tecnología Avanzada (RENATA), para: *el diseño, implementación y dinamización de la estrategia para el fortalecimiento de las Instituciones de Educación Superior en la producción, gestión y uso de Contenidos Educativos de Acceso Público*

		dirigidos a Educación Superior.
Año	**Acontecimiento**	**Descripción**
2011	Diseño de la Estrategia Nacional de Recursos Educativos Digitales Abiertos	Además de los explicado anteriormente, esta estrategia estima los procesos de producción, gestión y uso de los Contenidos Educativos, cifrados como Recursos Educativos Digitales Abiertos, que cumplan con tres condiciones para la consolidación de una oferta nacional de calidad: ser educativos, digitales y de acceso abierto.

Dentro de la historia de RI en Colombia para la difusión del conocimiento, se encuentran una serie de acontecimientos destacados en orden cronológico a partir del año 2005 y hasta el 2011 (ver Tabla 2.1), que van desde la creación del primer portal educativo de Colombia, Colombia Aprende, hasta el diseño de la estrategia nacional de recursos educativos digitales abiertos, pasando por otros acontecimientos como el proyecto de catalogación de OAs, la creación de la red nacional académica de tecnología avanzada y la creación de la biblioteca digital colombiana BDCOL, todos ellos encaminados a la promoción e instauración del acceso abierto como política educativa nacional.

2.1.4 Análisis de repositorios institucionales de Argentina y Colombia

Teniendo en cuenta el contexto histórico de los repositorios institucionales de Argentina y Colombia presentado en las secciones 2.1.1 y 2.1.2 respectivamente, en los Anexos 1 y 2 se muestran los resultados de la revisión de las principales características de los repositorios institucionales de acceso abierto más representativos de cada país.

Las características que se tuvieron en cuenta para este análisis comparativo fueron:

1. Plataforma sobre la cual está implementado el repositorio.

2. Estándar de metadatos utilizado.

3. Puesto que ocupa el repositorio en el Ranking Web de Repositorios Institucionales a nivel nacional, teniendo como referencia Argentina o Colombia, a nivel latinoamericano y a nivel mundial. Esta característica fue tenida en cuenta dado que el objetivo declarado del ranking es el de promover las iniciativas de acceso abierto, una de las formas más relevantes para la distribución de los resultados de investigación de las universidades y centros de investigación, a través del acceso gratuito a las publicaciones científicas en formato electrónico y otro tipo de materiales de carácter académico. Los indicadores web utilizados en este ranking (Consejo Superior de

Investigaciones Científicas, 2008) miden la visibilidad e impacto de los repositorios científicos.

El Ranking Web consiste en un listado de repositorios, que hospeden principalmente trabajos de investigación, clasificados de acuerdo a un indicador compuesto que combina datos de *presencia web* y de *impacto web* (visibilidad hipertextual), todo ello obtenido de los principales motores de búsqueda (Consejo Superior de Investigaciones Científicas, 2008).

4. Las opciones de búsqueda (general y avanzada), que ofrece el repositorio para localizar OA, y sus posibles combinaciones.

5. Las categorías, colecciones o tipos de OA que almacena el repositorio, en ocasiones con una breve descripción de los mismos, o referencias a páginas web donde se especifican en mayor nivel de detalle.

6. Las coincidencias y consistencias que se presentan en las búsquedas, es decir, los resultados de OA que se obtienen utilizando diversos criterios de búsqueda en las opciones que se presentan, realizando combinaciones entre los mismos o navegando a través de los enlaces relacionados con el listado de OA que se muestran al ejecutar una consulta.

7. En las observaciones se mencionan aspectos complementarios a la revisión, como son: referencia a la política de metadatos (cuáles de estos son obligatorios y cuáles opcionales), si se hace uso o no del sistema de autoarchivo, si el repositorio cuenta con un extractor automático de metadatos, si existe la posibilidad de exportar los resultados de las búsquedas en el formato del estándar de metadatos, entre otros. No todos estos aspectos se presentan para todos los repositorios, pues depende de la disponibilidad de información que se tenga al respecto.

La revisión se realizó navegando a través de cada una de las páginas web de los repositorios y recopilando información de interés para el estudio de la importancia y significado de los metadatos de OA dentro del proceso de búsqueda y su recuperación en un repositorio institucional, y de esta forma verificar cómo un adecuado manejo de los mismos hace posible una mejor descripción de los recursos educativos.

Las características sobre las cuales se prestó mayor atención fueron las opciones de búsqueda, coincidencias y consistencias en las búsquedas y categoría y colecciones que se trabajan. Esto se realizó teniendo en cuenta que son las que están más directamente relacionadas con el tema de metadatos en OA. Las demás características, consideradas como de apoyo o complementarias pero también asociadas con el tema de metadatos, se presentan o no, dependiendo de la disponibilidad de información que se tenga al respecto, tal como se mencionó anteriormente en el ítem "7. Observaciones".

A continuación se describe brevemente cada uno de los repositorios analizados, correspondiendo los primeros 7 a repositorios institucionales de Argentina (Anexo 1), mientras que los últimos 5 corresponden a repositorios institucionales de Colombia (Anexo 2.).

1) *SEDICI* (http://sedici.unlp.edu.ar/): Servicio de Difusión de la Creación Intelectual; es el Repositorio Institucional de la Universidad Nacional de la Plata, creado para albergar, preservar y dar visibilidad a las producciones de la Unidades Académicas de la Universidad. Como característica principal se destaca que es considerado el número 1 a nivel nacional y 5 a nivel Latinoamérica en el "Ranking Web de Repositorios del Mundo" (Consejo Superior de Investigaciones Científicas, 2008).

2) *Biblioteca Digital UNCuyo* (http://bdigital.uncu.edu.ar/): es el Repositorio Institucional de la Universidad Nacional de Cuyo, donde se da acceso libre a los cerca de 3740 objetos digitales publicados.

3) *Biblioteca Virtual UNL* (http://bibliotecavirtual.unl.edu.ar/): es un Repositorio Institucional de la producción científico-académica perteneciente a la Universidad Nacional del Litoral en formato digital; se divide en Bibliotecas que contienen colecciones específicas.

4) *Rephip – Repositorio Hipermedial UNR* (http://rephip.unr.edu.ar/): es un repositorio académico abierto creado para archivar, preservar y distribuir digitalmente en variados formatos tanto OA como la producción científica de Investigación y Desarrollo (I+D) de la comunidad académica de la Universidad Nacional de Rosario.

5) *Corciencia* (http://www.corciencia.org.ar/): repositorio digital de investigaciones científicas y tecnológicas que da acceso libre y abierto a la producción científica de la provincia de Córdoba.

6) *Biblioteca electrónica de ciencia y tecnología* (http://www.biblioteca.mincyt.gov.ar/index.php): es el portal argentino del conocimiento científico, desde el cual se tiene acceso a los artículos completos de más de 11.000 títulos de revistas científico-técnicas y más de 9.000 libros. Funciona en el marco de la Subsecretaría de Coordinación Institucional, dependiente de la Secretaría de Articulación Científico Tecnológica del Ministerio de Ciencia, Tecnología e Innovación Productiva.

7) *BDU² Repositorios Institucionales* (http://bdu.siu.edu.ar/cgi-bin/reporpt.pl): es un proyecto iniciando por el Consorcio SIU (Sistema de Información Universitario) para reunir recursos de información de valor académico de libre disponibilidad para el usuario final. Dichos contenidos pueden ser tesis de diversos grados académicos, artículos de publicaciones periódicas, libros electrónicos, material de alto valor histórico digitalizado, legislación educativa, videos, entrevistas y otro material puesto a disposición por instituciones académicas nacionales.

8) *Repositorio Institucional UN* (http://www.bdigital.unal.edu.co/): es el repositorio de Acceso Abierto de la Universidad Nacional de Colombia, en el cual se pretende administrar, preservar y difundir las obras monográficas que la Universidad ha producido a través de su historia, incluyendo libros, tesis y trabajos de grado, trabajos docentes, entre otros. Como característica principal se destaca que es considerado el número 1 a nivel nacional y 6 a nivel Latinoamérica en el "Ranking Web de Repositorios del Mundo" (Consejo Superior de Investigaciones Científicas, 2008).

9) *Repositorio Institucional EdocUR* (http://repository.urosario.edu.co/): en este repositorio se da acceso a textos completos de los documentos producidos por la Universidad del Rosario en su función docente, de investigación y de extensión.

10) *Repositorio Institucional PUJ* (http://repository.javeriana.edu.co/): es el repositorio de la producción intelectual de la Pontificia Universidad Javeriana.

11) *BDCOL- Biblioteca Digital Colombiana* (http://190.242.114.6/bdcol.html): es la Red Colombiana de Repositorios y Bibliotecas Digitales que indexa toda la producción académica, científica, cultural y social de las instituciones de educación superior, centros de investigación, centros de documentación y bibliotecas en general del país.

Allí se pueden encontrar alrededor de 85.000 documentos digitales en 73 Repositorios Institucionales de las diferentes regiones del país.

12) *Colombia Aprende* (http://aprende.colombiaaprende.edu.co/es/contenidoslo): este portal educativo, mencionado en la Tabla 1, surge como una iniciativa del Ministerio de Educación Nacional para elevar el nivel de educación en el país. Sus recursos están catalogados por asignatura, niveles de escolaridad, competencias, así como por el formato digital de los mismos.

De acuerdo a la información analizada para cada uno de estos repositorios, se pueden realizar las siguientes observaciones:

- 7 de los 12 repositorios revisados se encuentran entre los primeros 60 repositorios a nivel Latinoamérica dentro de 217 considerados en el "Ranking Web de Repositorios del Mundo" (Consejo Superior de Investigaciones Científicas, 2008), y dentro de los 10 primeros a nivel nacional (Argentina o Colombia, según corresponda), lo que da cuenta del gran trabajo y esfuerzo a nivel nacional en ambos países realizado hasta el momento en lo que a desarrollo repositorios institucionales se refiere.

- Los repositorios *Corciencia*, *Biblioteca electrónica de ciencia y tecnología*, *BDU*[2] por parte de Argentina, *BDCOL* y *Colombia Aprende* por parte de Colombia, no están considerados en el ranking por no pertenecer a una institución educativa como tal, sino albergar los recursos digitales científico-académicos y tecnológicos de un conjunto de instituciones.

- La mayoría de los repositorios revisados utilizan plataforma DSpace (DuraSpace, 2002) (software de código abierto para la implementación de un repositorio utilizado por más de 1.000 organismos e instituciones de todo el mundo para proporcionar un acceso sostenible a los recursos digitales) y estándar de metadatos DCMI (DCMI, Dublin Core Metadata Initiative, 1995), dando cuenta que este estándar es altamente usado en la dimensión educativa. Los demás usan plataformas como *EPrints* 3, EBSCO *DiscoveryService* y Protocolo Open Archives Initiative. Otros no mencionan la plataforma ni el estándar, siendo que probablemente no hagan uso de ninguno. Esto puede considerarse como una desventaja, en cuanto a la posibilidad de incluir un extractor automático de metadatos y a la interoperabilidad con otros repositorios.

- En cuanto a opciones de búsqueda inicial, la mayoría de los repositorios cuenta con una opción por palabra clave, o con agrupaciones por ciertas características particulares del material educativo, como por ejemplo, colecciones, autores o título, posiblemente definidos a través de los metadatos asociados a los OA.

- También existe la posibilidad de realizar búsquedas avanzadas, donde se pueden combinar diferentes criterios para acotar y detallar aún más los resultados de la búsqueda. Esto indica la gran importancia que tiene la información registrada en los metadatos asociados a los OA para que un usuario pueda tener éxito en encontrar lo que está buscando utilizando dichos criterios y filtros. De lo contrario, se pueden obtener recursos no esperados, o bien, pueden no encontrarse recursos que pueden ser de utilidad.

- La mayoría de los repositorios, exceptuando la Biblioteca Virtual de la UNL, utilizan el sistema de autoarchivo, lo que puede dar lugar a errores e inconsistencias en la carga y asociación de metadatos de los OA, debido a que está sujeto al criterio y conocimiento de los usuarios que almacenan los OA en el repositorio.

- En cuanto a categorías o tipos de colecciones, es muy diversa y variada la clasificación que trabajan los diversos repositorios. En general, están las de tipo texto (libros, publicaciones, artículos), video, audio, patente o marca, animación, multimedia e imágenes. Aquí también surge la importancia de contar con una adecuada clasificación del material educativo, para así poder definir y asociar a cada categoría los respectivos metadatos que lo describen de la mejor manera (Ver Anexo 3).

2.2 Metadatos y Estándares

Desde el punto de vista tecnológico, los metadatos constituyen una parte fundamental no sólo de los OAs para que éstos puedan ser encontrados y reutilizados, sino también de los repositorios en sí, ya que hacen posible que aumenten los niveles de confianza en la utilización de estas herramientas de búsqueda y consulta. A nivel educativo, los metadatos permiten evaluar la pertinencia y calidad de los OAs incluidos en los resultados de la búsqueda a partir de un fin educativo concreto, y de acuerdo al perfil del usuario que realiza la búsqueda.

Es posible encontrar dos tipos de metadatos en un OA: *objetivos* y *subjetivos*. Los valores de los primeros, al estar relacionados al contenido del OA, podrían ser asignados por medio de

software que utilice extracción automática. Los segundos, por el contrario, están asociados a información pragmática o intención de uso del OA y, por lo tanto, son datos provistos por los usuarios que archivan el objeto en los repositorios.

Tal como se muestra en la Tabla 2.2, los diferentes tipos de metadatos se pueden agrupar en 5 categorías de acuerdo a su funcionalidad, identificando algunos ejemplos de uso para cada una de las categorías. A nivel general se puede observar que los metadatos definidos para un OA sirven, por un lado, para apoyar la gestión y administración de los recursos de información -como es el caso de los metadatos *administrativos* y de *preservación*- y, por otro lado, para hacer posible el uso, integración y reutilización del OA dentro de un repositorio institucional -en referencia a los metadatos *descriptivos*, *técnicos* y de *uso*. En su conjunto, estos tipos de metadatos sirven para hacer una identificación integral de cualquier OA, logrando así el cumplimiento y estandarización de la mayoría de las características propias de los recursos de aprendizaje.

Tabla 2.2. Tipos de Metadatos (Baca, 1998).

Tipo de Metadatos	Descripción	Ejemplos de Uso
Administrativos	Metadatos usados para gestionar y administrar recursos de información	• Información sobre adquisición • Seguimiento de derechos • Documentación de requerimientos de acceso legal • Información sobre la posición • Criterios de selección • Control de versiones
Descriptivos	Metadatos usados para describir o identificar recursos de información	• Catálogos • Índices especializados • Relaciones entre recursos • Anotaciones hechas por usuarios
Preservación	Metadatos relacionados con la preservación de recursos	• Documentación de condiciones físicas de los recursos • Documentación de acciones llevadas a cabo para generar versiones
Técnicos	Metadatos relacionados con el funcionamiento de los sistemas	• Documentación sobre hardware y software • Información sobre formatos, razones de comprensión, rutinas de escalado, etc. • Passwords, llaves de encriptación
Uso	Metadatos relacionados con el nivel y tipología de uso de los recursos	• Muestra de registros • Seguimiento de uso y de usuarios • Reusado de contenido • Información sobre versiones

Teniendo en cuenta estos tipos de metadatos, se pueden identificar los dos grupos más importantes a nivel de OA. Por un lado los metadatos *descriptivos*, que abarcan aquellos atributos que están relacionados exclusivamente al OA y que lo pueden identificar de manera única. Por otro lado, los metadatos de uso, para hacer que el contenido del OA pueda ser reutilizado una y otra vez, sin restricciones.

La definición y adopción de estándares hace posible que los metadatos cumplan su función en mayor medida. Además, es recomendable para el almacenamiento y recuperación de los OA en los repositorios, ya que los metadatos son responsables de la interoperabilidad, aumentan las posibilidades de reutilización, eliminan las barreras tecnológicas y facilitan el análisis de calidad. Los estándares más utilizados y reconocidos a nivel de metadatos son DublinCore (DCMI, Dublin Core Metadata Initiative, 1995) e IEEE LOM (IEEE, 2002), y para OA en general, SCORM (Learning, 2000). A continuación se detallan cada uno de estos estándares.

2.2.1 DublinCore

DublinCore (DCMI, Dublin Core Metadata Initiative, 1995) es una organización abierta, iniciada en 1995, que está abocada al desarrollo de estándares de metadatos interoperables. El estándar Dublin Core (DCMI, DCMI Metadata Basics, 1995) se creó a partir de un taller de metadatos realizado precisamente en la ciudad de Dublin (Ohio), Estados Unidos. “La DCMI (Dublin Core Metadata Initiative) fue desarrollada para la descripción de un amplio universo de recursos en red, su aplicación es de carácter muy general e incluye los recursos que una biblioteca digital puede contener” (López Guzmán, 2005).

La primera versión generó grandes expectativas, pues se componía únicamente por un pequeño conjunto de descriptores con los cuales se podía describir en parte, y de forma muy sencilla, un recurso. En el año 2001 el organismo de normalización de los Estados Unidos aprobó como norma estatal el conjunto de elementos de DublinCore, dando lugar a la norma Z39-85:2001 DUBLIN CORE METADATA ELEMENT SET.

Con el auge de Internet y, principalmente por su nivel de inclusión en el mercado y en el ambiente académico, se comenzó a hacer un uso cada vez más extenso del estándar, razón por la cual en el 2003 se aprobó como norma ISO 15836, basándose ciertamente en la norma Z39-85.

El estándar DCMI (DCMI, DCMI Metadata Basics, 1995) cuenta con un conjunto de 15 definiciones semánticas que permiten la descripción y organización de la información, así

como también la definición de las propiedades de objetos para sistemas que se encarguen de la búsqueda de recursos basados en la Web. Los 15 elementos que componen el estándar son: *contribuidor*, *cobertura*, *creador*, *fecha*, *descripción*, *formato*, *identificador*, *lenguaje*, *editor*, *relación*, *derechos*, *fuente*, *tema*, *título* y *tipo* (Figura 2.1).

A su vez, estos se agrupan en 3 grandes categorías: **contenido**, **propiedad intelectual** e **instanciación.** La primera de estas categorías, ***contenido***, agrupa los metadatos utilizados para describir el contenido del recurso, como por ejemplo el lenguaje en el que se encuentra escrito, el ámbito en el que puede ser usado, un resumen que explica el contenido, su título entre otras. La segunda categoría, ***propiedad intelectual***, agrupa los metadatos utilizados para identificar al autor del recurso y sus derechos. Finalmente, la categoría ***instanciación***, agrupa los metadatos utilizados para identificar a esa instancia propiamente dicha, en caso en que puedan existir réplicas del recurso. Metadatos de esta categoría son la fecha en la que el recurso fue publicado o se encuentra disponible, el tipo de recurso, el formato físico y una referencia unívoca al mismo.

En el DublinCoreMetadataElement Set (versión 1) se describe cada una de las 3 categorías mencionadas anteriormente. La Tabla 2.3 muestra cómo se clasifican cada uno de los 15 elementos del estándar dentro de las categorías y, adicionalmente, cuenta con una etiqueta, un descriptor y una definición que en su conjunto permiten identificarlo dentro de un OA.

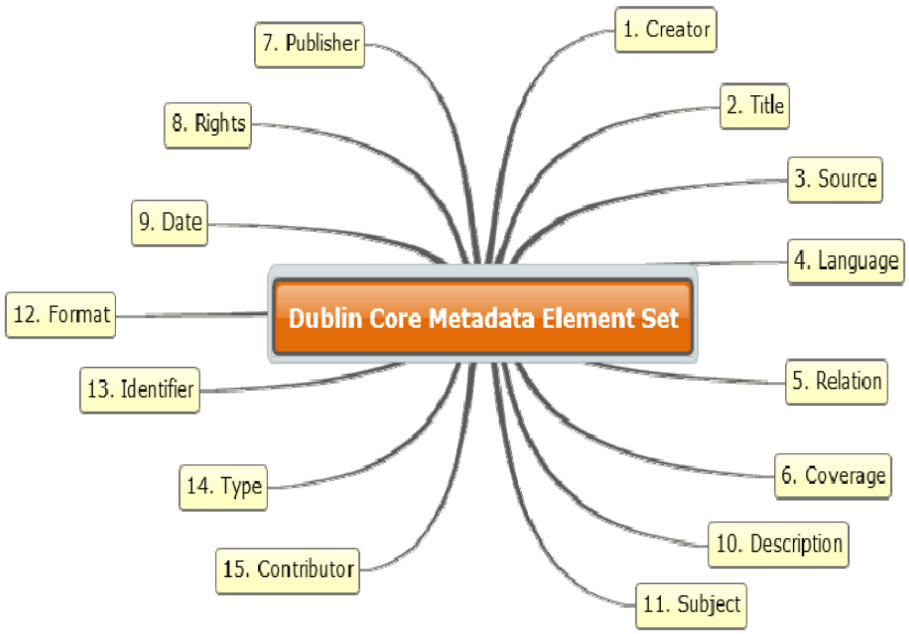

Fig. 2.1. Jerarquía completa de DCMI. Tomado de Figura 4.2 (Astudillo, 2011)

Cada uno de estos descriptores lleva consigo una serie de atributos, que son comunes a todos ellos, tales como: la versión, el registro de autoridad DCMI, el lenguaje, la obligatoriedad, el tipo de dato que maneja y el número máximo de ocurrencias dentro del OA, que pueden guiar su implementación (ISO/IEC, 2003).

DCMI cuenta también con una serie de *cualificadores* que redefinen de cierta manera cada uno de los descriptores que componen el estándar, manteniendo su esencia, refinando la descripción del recurso que se está catalogando, y aumentando la especificidad y precisión del metadato que sea asignado a dichos recursos.

Tabla 2.3 Definición de elementos de DublinCore.

Etiqueta	Descriptor	Definición
1. Contenido		
dc.title	Title/Título	Título o nombre dado al recurso.
dc.source	Source/Fuente	Si fuera necesario, se puede describir el recurso desde donde fue derivado. Cadena de texto que describe de forma única al recurso o el sitio donde se encuentra el recurso.
dc.language	Language/Lenguaje	Idioma del recurso.
dc.relation	Relation/Relación	Un recurso con el que el material se puede relacionar. Enlace a un segundo o tercer recurso que se relaciona con el recurso original. Al igual que el descriptor *Source* es un enlace único e inequívoco del recurso en el sistema.
dc.coverage	Coverage/Cobertura	Característica de cobertura espacial y/o temporal del contenido intelectual del recurso. Dónde el recurso es aplicable, o la jurisdicción en la cuál es relevante.
dc.description	Description/Descripción	Descripción o resumen del recurso. Puede incluir tabla de contenidos o una representación gráfica del recurso.
dc.subject	Subject/Tema	El tema sobre el que trata el recurso. También puede llevar información sobre palabras clave asociadas.
2. Propiedad Intelectual		
dc.creator	Creator/Creador	Entidad o persona encargada de la creación del recurso.
dc.publisher	Publisher/Editor	Entidad o persona responsable de publicar el recurso.
dc.rights	Rights/Derechos	Información sobre los derechos de autor del recurso, para que el usuario final conozca sus condiciones de uso y acceso.
dc.contributor	Contributor/Contribuidor	La entidad responsable de hacer contribuciones a los recursos.
3. Instanciación		
dc.Date	Date/Fecha	Punto o un período de tiempo asociado con un evento en el ciclo de vida del recurso; por ejemplo, fecha en la cual se creó o se publicó el recurso.
dc.format	Format/Formato	Formato de archivo, medio físico, o el espacio de almacenamiento o duración del recurso.
dc.identifier	Identifier/Identificador	Referencia no ambigua para el recurso en un contexto dado, como por ejemplo, ISSN, ISBN, etc.
dc.type	Type/Tipo	La naturaleza o género del recurso, como por ejemplo, tesis, libros, artículos, revista, etc.

Por lo tanto, este estándar de metadatos no está restringido a un perfil de aplicación específico, y es altamente usado en el mundo en diferentes disciplinas de estudio. Muchos repositorios lo han adoptado para etiquetar sus recursos de material educativo (por ejemplo, SEDICI, Rephip, Corciencia, Universidad Nacional de Colombia, Universidad Javeriana referenciados en la sección 2.1).

DCMI puede ser utilizado sobre cualquier sistema de información y, a su vez, permite que dicho sistema sea interoperable con otros sistemas de información que ofrezcan sus contenidos según las etiquetas definidas por DublinCore.

Finalmente, es importante señalar que, según la iniciativa DublinCore, la relación entre un registro de metadatos y el recurso que se desea describir debe darse por alguna de estas dos formas:

1) Los elementos pueden estar en un registro separado del documento, como en el caso del registro de un catálogo de bibliotecas, o
2) los metadatos pueden estar incluidos, incrustados, en el propio recurso.

Teniendo en cuenta esta relación, a nivel de diseño de extractores automáticos, es más simple y útil que los metadatos se encuentren en archivos separados del OA. Según (López Guzmán, 2005) esta manera de tener los metadatos separados del recurso, facilita la indexación y la reutilización de dichos metadatos, logrando mayor independencia.

2.2.2 IEEE LOM

El IEEE LTSC (IEEE Learning Technology Standards Committee) trabaja para el desarrollo y mantenimiento de un estándar de metadatos para OA desde 1997 denominado Learning Object Metadata (LOM) (IEEE, 2002). Este estándar es el fruto de un esfuerzo internacional del LOM WorkingGroup (o WG12), con miembros que representan a más de 15 países. En Junio de 2002, la IEEE LTSC completa y publica el 1484.12.1 LOM data model standard (Barkman, y otros, 2002). LOM es uno de los primeros estándares de metadatos que fue diseñado específicamente para describir material educativo, en particular OA, y es uno de los más difundidos y utilizados.

El modelo especifica cómo deberían ser descriptos los OA. Cuenta con nueve categorías: *general*, *ciclo de vida*, *meta-metadatos*, *técnico*, *enseñanza*, *derechos*, *relación*, *anotación* y *clasificación*. Las categorías, a su vez, contienen subcategorías. El modelo cuenta con un total

de 76 elementos o campos para rellenar, que además son extensibles. La jerarquía completa de IEEE LOM se puede ver en la Figura 2.3.

El estándar IEEE 1484.12.1:2002 sobre metadatos para Objetos de Aprendizaje describe cada categoría como sigue (Barkman, y otros, 2002):

- General. Información general que describe un objeto educativo.
- Ciclo de vida. Características relacionadas con la historia y el estado actual del objeto, y aquellas que lo han afectado durante su evolución.
- Meta-metadatos. Información sobre la propia instancia de metadatos.
- Técnica. Requerimientos y características técnicas del objeto.
- Uso educativo. Características educativas y pedagógicas del objeto.
- Derechos. Derechos de propiedad intelectual y condiciones para el uso del objeto.
- Relación. Características que definen la relación entre el objeto y otros objetos relacionados.
- Anotación. Permite incluir comentarios sobre el uso educativo del objeto e información sobre cuándo y por quién/es fueron creados los comentarios.
- Clasificación. Describe el objeto en relación a un determinado sistema de clasificación.

De igual manera, la estructura de los metadatos LOM se puede representar en una jerarquía en árbol, como por ejemplo el de la Figura 2.2 que representa los elementos que pertenecen a la categoría "General". El nodo raíz corresponde al recurso que se está describiendo y suele recibir el nombre de *lom*. En el siguiente nivel están los subelementos o elementos intermedios, llamados ramas, que pueden contener, a la vez, otros subelementos o elementos terminales, llamados hojas. Para cada elemento en la jerarquía se especifican la definición, tipo de datos, valores permitidos y si permite o no multiplicidad.

Es importante mencionar que se han propuesto equivalencias que permiten pasar de DCMI a IEEE LOM, como se muestra en la Tabla 2.4, donde los 15 elementos definidos en DCMI se corresponden de forma directa con elementos definidos en el estándar IEEE LOM.

También existen modelos basados en los dos estándares, como MIMETA (Marzal García-Quismondo, Calzada Prado, & Cuevas Cerveró, 2006)(Modelo IACORIE de Metadatos), que busca unir la simplicidad de DCMI con la riqueza descriptiva de IEEE LOM y, a su vez, permite mantener una interoperabilidad entre repositorios que usan dichos estándares en la descripción de sus recursos.

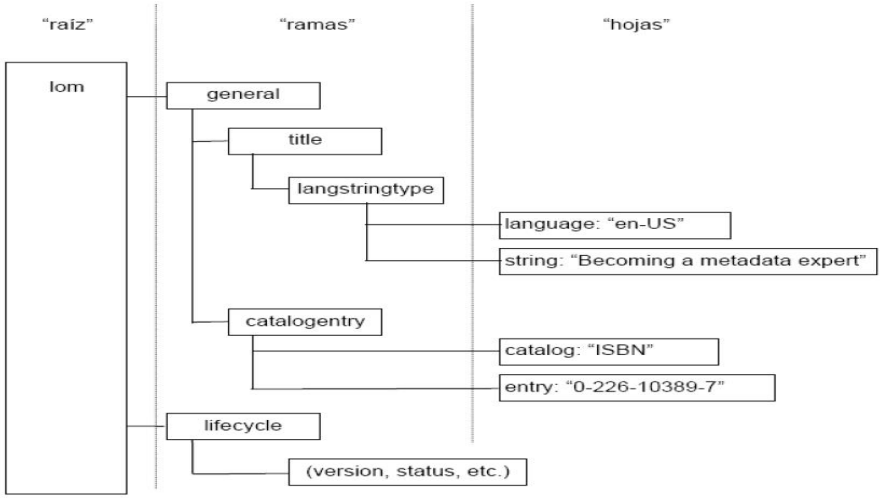

Fig. 2.2. Jerarquía en árbol de IEEE LOM. Tomado de Figura 5 (Blanco Suárez, 2006)

Otra propuesta interesante es el compromiso existente del grupo de trabajo LOM con la Iniciativa de Metadatos de Dublin Core en el desarrollo de metadatos interoperables para el aprendizaje, la educación y la formación, tal como se refleja en el Memorando de Entendimiento entre el IEEE LTSC LOM WG y el DCMI (DCMI-IEEE-MOU, 2000). Con este compromiso, ambas organizaciones se ven beneficiadas a través de la unión de esfuerzos en la utilización de herramientas y servicios que emergen de las dos iniciativas, mediante el uso de una arquitectura única que presenta muy pocas barreras para la creación, intercambio y uso de metadatos estructurados.

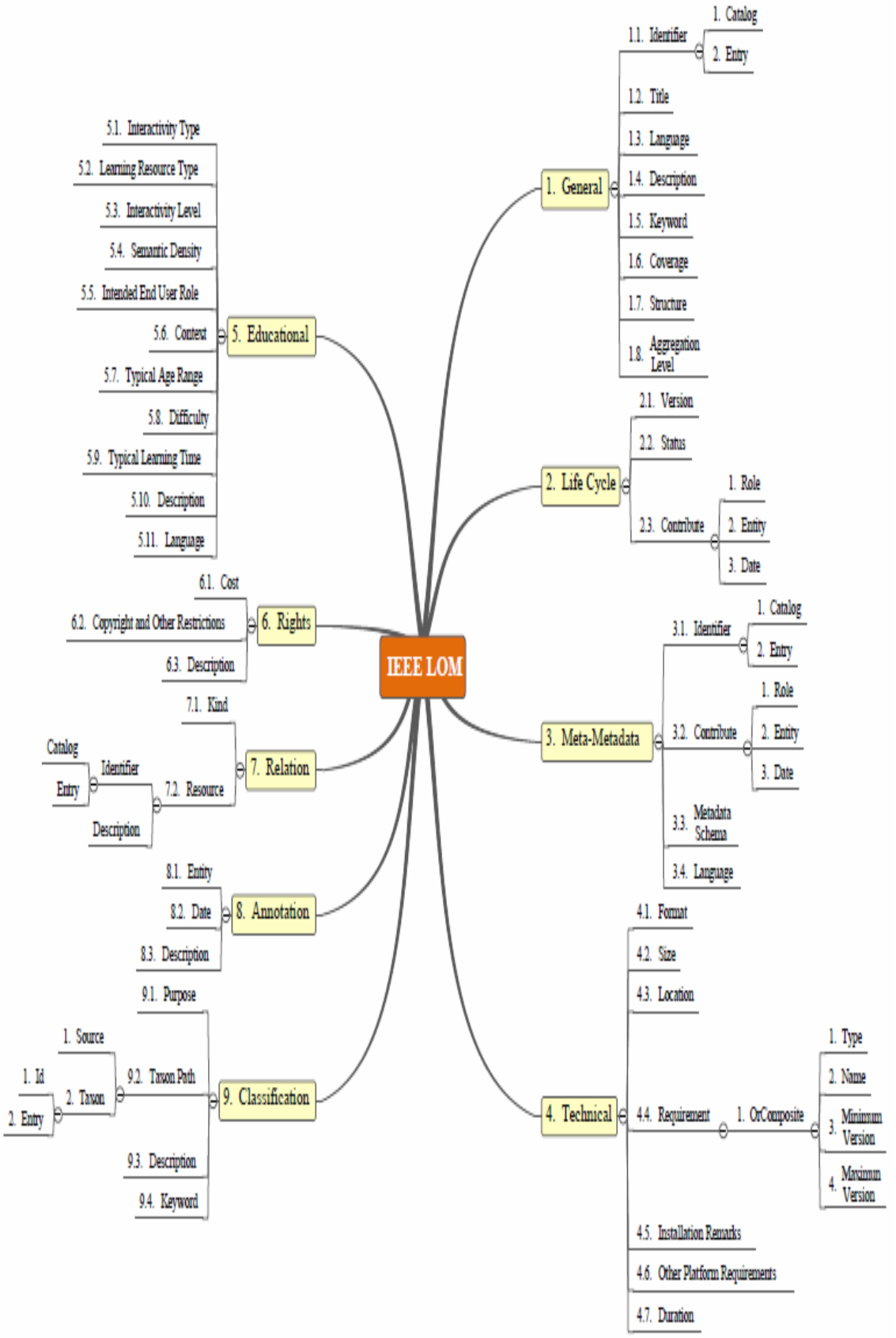

Fig. 2.3. Jerarquía completa de IEEE LOM. Tomada de Figura 4.1 (Astudillo, 2011)

Tabla 2.4. Tabla de equivalencias de elementos DCMI e IEEE LOM (Barkman, y otros, 2002).

Campo	Metadato DCMI	Descripción del Campo	Equivalente IEEE LOM
Identificador del Recurso	DC.Identifier	Secuencia de caracteres usados para identificar unívocamente un recurso. Por ejemplo la URL del recurso.	1.1.2:General.Identifier.Entry
Título	DC.Title	El nombre dado a un recurso, normalmente proporcionado por el autor.	1.2:General.Title
Lengua	DC.Language	Lengua/s del contenido intelectual del recurso	1.3:General.Language
Descripción	DC.Description	Una descripción textual del recurso, como puede ser un resumen o una descripción de su contenido.	1.4:General.Description
Claves	DC.Subject	Expresa las claves o frases que describen el título o el contenido del recurso.	1.5:General.Keyword o 9:Classification con 9.1:Classification.Purpose igual a "disciplina" o "idea".
Cobertura	DC.Coverage	La característica de cobertura espacial y/o temporal del contenido intelectual del recurso.	1.6:General.Coverage
Tipo de Recurso	DC.Type	La categoría del recurso.	5.2:Educational.Learning Resource Type
Fecha	DC.Date	Una fecha en la que el recurso se puso a disposición del usuario en su forma actual.	2.3.3:Life Cycle.Contribute.Date cuando 2.3.1:Life Cycle.Contribute.Role tiene como valor "editor".
Autor	DC.Creator	La persona u organización responsable de la creación del contenido intelectual del recurso.	2.3.2:Life Cycle.Contribute.Entity cuando 2.3.1:Life Cycle.Contribute.Role tiene como valor "autor".
Otros Colaboradores	DC.OtherContributor	La persona u organización que haya tenido una contribución intelectual significativa en la creación del recurso, pero cuyas contribuciones son secundarias en comparación a las de las personas u organizaciones especificadas en el elemento *Creator.*	2.3.2:Life Cycle.Contribute.Entity con el tipo de contribución especificada en 2.3.1:Life Cycle.Contribute.Role
Editor	DC.Publisher	La entidad responsable de hacer que el recurso se encuentre disponible en la red en su formato actual.	2.3.2:Life Cycle.Contribute.Entity cuando 2.3.1:Life Cycle.Contribute.Role tiene como valor "editor".
Formato	DC.Format	El formato de datos y/o las dimensiones (tamaño, duración…) de un recurso usado para identificar el software y posiblemente, el hardware que se necesitaría para mostrar el recurso.	4.1:Technical.Format

Campo	Metadato DCMI	Descripción del Campo	Equivalente IEEE LOM
Derechos	DC.Rights	Una referencia (URL, por ejemplo) para una nota sobre derechos de autor, para un servicio de gestión de derechos o para un servicio que dará información sobre términos y condiciones de acceso a un recurso.	6.3:Rights.Description
Relación	DC.Relation	Un identificador de un segundo recurso y su relación con el recurso actual. Este elemento permite enlazar los recursos relacionados y las descripciones de los recursos.	7.2.2:Relation.Resource.Description
Fuente	DC.Source	Secuencia de caracteres utilizado para identificar unívocamente un trabajo a partir del cual proviene el recurso actual.	7.2:Relation.Resource cuando el valor de 7.1:Relation.Kind es "sebasaen".

2.2.3 SCORM

El Departamento de Defensa de los Estados Unidos a través de ADL (Advanced Distribuided Learning) desarrolló un modelo denominado SCORM (Shareable Content Object Reference Model) (Learning, 2000), a partir de un conjunto de estándares y especificaciones interrelacionadas. Este estándar se construyó en base al trabajo de otras organizaciones de estándares, como son AICC, IMS, IEEE LTS y ARIADNE, con la finalidad de crear un modelo de contenidos para el aprendizaje centrado en la Web.

La utilización de SCORM permite el empaquetamiento del contenido, actividades y metadatos, propiciando la accesibilidad, reutilización y durabilidad, y facilitando la migración de OA entre diferentes ambientes virtuales de aprendizaje que hagan uso del estándar.

El contenido del estándar se encuentra publicado en un libro resumen y tres libros técnicos, tal como se muestra de forma ilustrativa en la Figura 2.4.

- The Content AggregationModel (CAM), *modelo de agregación de contenidos*: este libro describe los tipos de objetos utilizados en una agregación de contenidos, cómo empaquetar los mismos, cómo describir los contenidos a través de metadatos bajo el uso de algún estándar como los vistos anteriormente, y cómo definir la secuenciación de contenidos. Dentro del paquete de contenidos se encuentra el “manifiesto” o

esquema conceptual (Figura 2.5), como base para la comunicación con el ambiente virtual de aprendizaje.

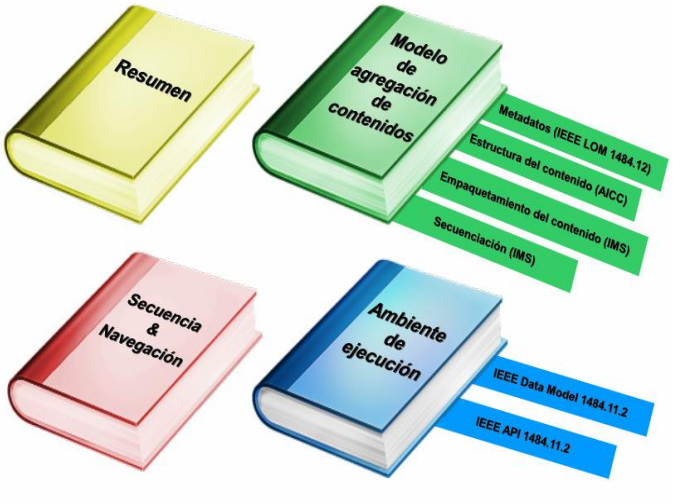

Fig. 2.4. Biblioteca SCORM.

- The Run-Time Environment (RTE), *ambiente de ejecución*: este libro describe los requerimientos necesarios para los ambientes virtuales de aprendizaje, con el fin de permitir la interoperabilidad entre diferentes sistemas.
- Sequencing&Navigation (SN), *secuenciación & navegación*: este libro define el método para representar el comportamiento natural previsto para la secuencia de aprendizaje, de forma tal que cualquier ambiente virtual de aprendizaje que haga uso del estándar, ejecute el paquete SCORM consistentemente.

Varios ambientes virtuales de aprendizaje como Moodle, Dokeos, Ilias, e-ducativa, Blackboard, entre otros, han adoptado el uso del estándar SCORM, integrando así en sus OAs la norma y uso de los metadatos, el empaquetamiento y secuenciamiento.

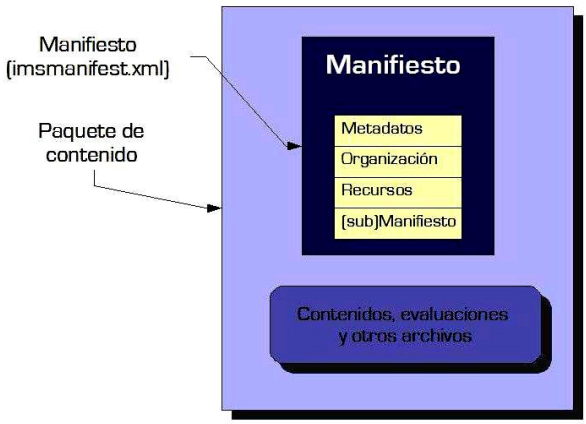

Fig. 2.5. Esquema conceptual de un paquete de contenidos.

2.3 Técnicas de Procesamiento de Lenguaje Natural (NLP)

El lenguaje natural es aquel que utilizan los seres humanos para comunicarse unos con otros. A su vez, el procesamiento del lenguaje natural (PLN o NLP por sus siglas en inglés *Natural Language Processing*) se encarga del procesamiento computacional del lenguaje natural y de cómo aplicarlo para dar solución a problemas de ingeniería. El procesamiento del lenguaje natural involucra una transformación a una representación formal, manipula esta representación y por último, si es necesario, lleva los resultados nuevamente a lenguaje natural (Hernández & Gómez, 2013).

El origen de esta técnica se da a finales de la década de 1940. En sus inicios, durante la década de 1950, se encontraron muchas limitaciones, como la existencia de modelos morfológicos y sintácticos poco evolucionados, y el poco interés en la comprensión del significado. Ya para la década de 1960 se dio un cambio de enfoque hacia el procesamiento de frases y comprensión, además del desarrollo y uso de una interfaz amigable, así como también el desarrollo de formalismos y técnicas de análisis sintáctico.

Entre los años 70 y 80, se realizaron nuevos desarrollos basados en formalismos y aumentaron los campos de aplicación de las técnicas NLP, centrando el interés en la representación del significado, hacia la comprensión de lenguaje, interfaz a base de datos, enseñanza asistida por ordenador, automatización de tareas de oficina, programación automática y procesamiento de texto científico, entre otras.

En los últimos años, las aplicaciones en el área llegan a los usuarios finales y han dado lugar a varias líneas de investigación como son:

- Revisión lingüística de textos
- Recuperación y extracción de información
- Traducción automática
- Generación automática de resúmenes de texto
- Reconocimiento de entidades y conceptos
- Comprensión y formalización del lenguaje natural
- Reconocimiento óptico de caracteres (OCR por sus siglas en inglés)
- Reconocimiento y síntesis de voz
- Etiquetado morfológico, semántico y sintáctico
- Análisis de sentimientos del texto

Los principales objetivos de las técnicas de procesamiento de lenguaje natural se pueden resumir como sigue:

1) Simplificar la comunicación hombre-máquina (HCI, Human-Computer Interaction): a través de una nueva dimensión en las aplicaciones informáticas que permita expresar órdenes en lenguaje natural y que dichas órdenes sean entendidas y ejecutadas por las máquinas.
2) Procesamiento y análisis de textos: las necesidades informáticas de los usuarios finales son cada vez mayores; es por esto que las técnicas NLP contribuyen significativamente en el análisis de textos para la extracción automática de conocimiento y recuperación de información relevante, que luego es utilizada para generar más conocimiento.
3) Traducción automática: también llamada MT (Machine Translation por sus siglas en inglés). El nivel básico consiste en la traducción literal del texto de un lenguaje natural a otro; un nivel más avanzado implica traducciones más complejas, reconocimiento de frases, traducción de expresiones idiomáticas y un manejo más adecuado de las similitudes gramaticales.

La principal dificultad de las NLP reside en la ambigüedad del lenguaje natural que debe ser tenida en cuenta a la hora de analizar un texto: los diferentes significados de las palabras (polisemia), diferentes recursos literarios, regionalismos, errores ortográficos, sentimiento o carga emocional de las oraciones. La mayoría de las veces estas ambigüedades son difíciles de identificar incluso para quien está leyendo un texto, lo que complica su representación en un lenguaje formal y aún más su interpretación por parte de una máquina.

Las NLP tienen cuatro componentes básicos de análisis que facilitan y optimizan las tareas de procesamiento y extracción de la información. La ejecución de estos componentes suele ser secuencial, aunque también es posible la ejecución simultánea, y en el orden que se presentan a continuación:

1) *Análisis léxico o morfológico*

   Una primera parte consiste en analizar y dividir el texto en una serie de componentes léxicos o tokens, pertenecientes al lenguaje en el que está escrito el texto, considerando especialmente los separadores léxicos definidos para el lenguaje en cuestión. Por ejemplo, en los lenguajes naturales los separadores léxicos corresponden a los espacios en blanco y

los signos de puntuación. La construcción de analizadores léxicos está basada en los Autómatas Finitos Deterministas (AFD) (Vilca, 2014)

Otra parte hace referencia al etiquetado morfológico POS Tagging (Part-of-speech tagging) que se explica más adelante en la sección 2.3.4

2) *Análisis sintáctico*

Determina si una secuencia de componentes léxicos o tokens cumplen una determinada estructura gramatical, siguiendo las reglas gramaticales del lenguaje analizado. Esto se logra haciendo uso de los resultados obtenidos del análisis léxico y el etiquetado morfológico.

3) *Análisis semántico*

Determina si una secuencia de componentes léxicos o tokens forman una sentencia bien construida, coherente y con sentido, haciendo uso del árbol sintáctico generado por el analizador sintáctico.

4) *Análisis pragmático*

Analiza la relación entre las palabras y el contexto donde son utilizadas, evaluando la influencia del mismo en su significado e interpretación. Este análisis se lleva a cabo haciendo uso de los resultados del análisis semántico.

A continuación se presentarán cinco técnicas de procesamiento de lenguaje natural que fueron tenidas en cuenta durante el diseño y desarrollo del algoritmo AMELOIR:

2.3.1 Expresiones Regulares

Una expresión regular sirve como un descriptor de un lenguaje. También es una herramienta para describir patrones de texto (Vilca, 2014). Formalmente, el objetivo de las expresiones regulares es representar todos los posibles lenguajes definidos sobre un alfabeto ∑, en base a una serie de lenguajes primitivos, y operadores de composición (Billhardt, 2007).

Los lenguajes primitivos son: el lenguaje vacío, el lenguaje formado por la palabra vacía, y los lenguajes correspondientes a los distintos símbolos del alfabeto. Los operadores de composición son la unión, la concatenación y el cierre.

La definición de las expresiones regulares se da de la siguiente manera: dado un alfabeto ∑, las expresiones regulares sobre ∑ se definen de forma recursiva por las siguientes reglas:

1. Las siguientes expresiones son expresiones regulares primitivas:

- ✓ ϕ
- ✓ λ
- ✓ a, siendo $a \in \Sigma$

2. Sean $\alpha\ y\ \beta$ expresiones regulares, entonces son expresiones regulares derivadas:

- ✓ $\alpha + \beta$ denota la **unión** de los lenguajes que representan dichas expresiones regulares.
- ✓ $\alpha \cdot \beta$ (o simplemente $\alpha\beta$) denota la **concatenación**.
- ✓ α^* denota la **estrella de Kleene** o **clausura de Kleene**.
- ✓ (α) denota el mismo lenguaje α.

3. No hay más expresiones regulares sobre ∑ que las construidas mediante estas reglas.

Dentro de las propiedades de expresiones regulares, se encuentra la definición de equivalencia: dos expresiones regulares $r_1\ y\ r_2$ se dicen *equivalentes*, $r_1 = r_2$, si describen el mismo lenguaje, esto es, si $L(r_1) = L(r_2)$.

Así mismo, dentro de las aplicaciones se encuentran:

1. Se utiliza una notación similar a una expresión regular para describir patrones de búsqueda. Por ejemplo, en el explorador de archivos se pueden usar **.doc* para indicar todos los archivos que tengan cualquier nombre y cuya extensión sea "doc".

2. En los generadores de analizadores léxico como el Lex y el Flex. El análisis léxico es una fase de un compilador que divide el código de entrada en unidades lógicas denominadas tokens de uno o más caracteres. Ejemplos de tokens son las palabras reservadas (como *if, while, for,* entre otros), identificadores y signos como +,>.

3. En búsqueda de texto flexible, se utiliza la expresión regular como un patrón de búsqueda; por ejemplo, *a??bc[f-p]* puede representar a todas la cadenas (palabras) de texto que inician con la *a*, seguida de cualquiera dos caracteres (símbolos), seguida de *bc*, y, finalmente, un carácter que puede ser desde la *f* hasta la *p*, es decir, uno del rango de caracteres comprendidos entre la *f* y *p*.

4. Sistemas de formateo de texto que usan notación de tipo expresión regular para describir patrones.

Una ventaja adicional de las expresiones regulares es que se pueden convertir en una máquina (autómata finito), el cual puede automáticamente decidir si una cadena (palabra) pertenece o no al lenguaje denotado por la expresión regular, es decir, ofrecen una manera declarativa de expresar las cadenas que se aceptan en el lenguaje.

2.3.2 Stop Words

Las Stop Words o palabras vacías tienen una función gramatical, pero no aportan nada al contenido del documento, es decir, tienen un rol más sintáctico que semántico. Mediante la eliminación de tales términos se puede contribuir a mejorar la recuperación de la información, ya que es poco probable que los usuarios realicen búsqueda de documentos por esas palabras.

Los estudios correspondientes a este fenómeno fueron iniciados por Hans Peter Luhn en 1958 con su investigación sobre el índice KWIC (Keyword in Context) (Association for Information Science and Technology, 2014), una técnica de indexación que organiza las palabras según su consideración como claves para la recuperación o no de la información, teniendo en cuenta el contexto del documento. Este proceso derivó en la acuñación del término "palabra vacía" para referirse a aquellas con un bajo poder discriminatorio y representativo del contenido del documento. Los análisis estadísticos efectuados por Luhn demostraron que la indexación es un proceso más rápido cuando se prescinde de tales términos, favoreciendo el ahorro de espacio requerido para el almacenamiento de la información. También se demostró que entre un 30% y un 50% de las palabras de un texto corresponden a tal categoría.

Una de las técnicas más utilizadas para remover palabras vacías consiste en revisar si la palabra está dentro de un listado que contiene las palabras vacías más comunes, como son adverbios, conjunciones, preposiciones, entre otras. Sin embargo, cada colección de documentos es única. Por lo tanto, es razonable tener un listado de palabras vacías diferente para diferentes colecciones, con el fin de maximizar el rendimiento de un algoritmo de recuperación de la información.

No obstante, la técnica de eliminación de palabras vacías ha mejorado, debido a la introducción de técnicas que tienen en cuenta el significado de tales palabras cuando están acompañadas de sustantivos, en casos en los que no pueden ser separadas o eliminadas por

conformar una denominación propia, así como por pérdidas en el significado semántico de un sintagma, frase o palabra (Blázquez Ochando, 2013).

2.3.3 Stemming

El proceso stemming o lematización, consiste en la remoción de los sufijos en palabras que pertenezcan a la misma familia semántica, es decir, reducir las palabras a sus elementos mínimos con significado, las raíces de las palabras.

El objetivo tras remover sufijos de palabras con significados similares es reducir la cantidad de dispersión de los conceptos presentes en un documento, acotando las terminaciones de las palabras a su forma más genérica o común y, por consiguiente, aumentar la precisión de los algoritmos de recuperación de la información que se utilicen para extraer conocimiento del documento en cuestión. La implementación de estos algoritmos están vinculados a los idiomas y por esto es que se encuentran más implementaciones para palabras en inglés y muy pocas para el español. Por lo que, tanto en inglés como en español y en cualquier idioma, un término puede ser reducido a su común denominador, permitiendo la recuperación de todos los documentos cuyas palabras tengan la misma raíz común, como por ejemplo: *catálogo, catálogos, catalogación, catalogador, catalogar, catalogando, catalogado, catalogándonos*. Todos los términos derivan en tal caso de "*catalog*", haciendo posible que la recuperación sea completa en más de 8 supuestos distintos. No obstante, no siempre esta técnica permite resolver perfectamente todas las consultas que un usuario pueda plantear

2.3.4 POS Tagging

POS Tagging (Part-of-speech tagging) resuelve el problema de asignar a cada palabra de una frase, la función que cumple en esa frase, es decir, se trata de un etiquetado gramatical.

La entrada para el proceso de POS Tagging es un texto más un conjunto de etiquetas, y la salida del proceso es un texto etiquetado, asignando a cada palabra una única etiqueta. Esta sencilla técnica, puede ser llevada a cabo por diferentes métodos que tienen en cuenta el contexto local, y es utilizada en muchas aplicaciones, como por ejemplo, reconocimiento de voz, análisis sintáctico, recuperación de información, entre otras.

2.3.5 Named Entity Recognition (NER)

El reconocimiento de entidades nombradas (NER por sus siglas en inglés), también conocido como extracción de entidades, permite localizar y clasificar entidades dentro de un texto, de

acuerdo a categorías tales como nombres de personas, organizaciones, ubicaciones, expresiones de tiempo, cantidades, valores monetarios, porcentajes, entre otros. Se han creado sistemas NER que utilizan técnicas basadas en gramática lingüística, así como modelos estadísticos, es decir, aprendizaje automático.

El reconocimiento de entidades nombradas a menudo se divide, conceptualmente y posiblemente también en su implementación, en dos problemas distintos: detección de nombres y clasificación de los nombres según el tipo de entidad al que hacen referencia (persona, organización, ubicación y otro).

Capítulo

3

Extractores de Metadatos: estado del arte

Anteriormente se mencionó la importancia de la extracción automática de metadatos como una forma de garantizar la calidad de la información de los OAs que se almacenan en los repositorios, y que será utilizada durante la ejecución de las búsquedas de material educativo. Esto también, de alguna manera, asistirá a los usuarios en la selección de OAs relevantes a sus preferencias y necesidades. Lo anterior, sumado a la estandarización de metadatos que se presentó en la Sección 2.2 y el auge de los repositorios institucionales y del acceso abierto al conocimiento que se vislumbró y describió en la Sección 2.1, da como resultado el fundamento necesario para comprender la verdadera importancia del desarrollo de nuevos algoritmos para la extracción automática de metadatos a partir de OA en RI.

Existen tres aspectos importantes que se deben tener en cuenta al momento de elegir o diseñar un sistema de este tipo:

a) Los tipos de archivos a procesar (por ejemplo, html, txt, pdf, doc, etc.),

b) Los metadatos que son extraídos, que conforman uno de los puntos más importantes a ser considerado, poniendo especial atención en aquellos a nivel educativo, y

c) Las técnicas y recursos utilizados para realizar la extracción, por ejemplo, herramientas para el procesamiento de lenguaje natural (NLP), ontologías, etc.

A la luz de estos tres aspectos, se revisaron las siguientes siete propuestas existentes para la extracción automática de metadatos en OA, con el fin de tener en cuenta aspectos relevantes en el diseño del nuevo algoritmo:

1) SAXEF (System for Automatic eXtraction of E-learning object Features) (Alfano, Lenzitti, & Visalli, 2007): es un sistema creado por The Center on Communication Studies

(Universidad de Palermo, Italia); extrae indicadores didácticos de páginas Web en forma automática.

2) TWYS (Tang Way Yuen System) (Wai Yuen, 2007): este sistema fue desarrollado por Tang Way Yuen en su Tesis de Magister en Filosofía, dentro del departamento de Ciencias de la Computación en la Ciudad Universitaria de Hong Kong. Realiza la extracción automática de metadatos de páginas web HTML, bajo el estándar LOM.

3) LookIng4LO (Sistema Informático para la Extracción Automática de Objetos de Aprendizaje) (Motz, y otros, 2009): fue creado en el Instituto de Computación de la Facultad de Ingeniería (Universidad de la República, Uruguay). Es un sistema genérico y flexible, capaz de extraer OAs con sus respectivos metadatos de archivos XML y HTML, documentos Word, presentaciones PowerPoint, archivos PDF y paquetes SCORM.

4) MAGIC (Metadata Automated Generation for Instructional Content) (Li, Dorai, & Farrell, 2005): sistema desarrollado en el centro de investigación IBM Watson, que automáticamente identifica, segmenta y genera metadatos críticos de acuerdo al estándar SCORM para contenidos educacionales.

5) Asistente para el depósito de objetos en repositorios con extracción automática de metadatos (Casali, Deco, Bender, Fontanarrosa, & Sabater, 2013): implementado en el Repositorio Hipermedial Institucional de la Universidad Nacional de Rosario, Argentina, en el año 2013. Propone modificar el flujo de carga estándar de la plataforma DSpace y diseñar un asistente de carga para la extracción automática de algunos metadatos, con el fin de incorporarlo a un RI.

6) CERMINE (Content ExtRactor and MINEr) (Tkaczyk, Szostek, Dendek, Fedoryszak, & Bolikowski, CERMINE - Automatic extraction of metadata and references from scientific literature, 2014): es un flujo de trabajo para el proceso de extracción de metadatos en documentos de origen digital, desarrollado en el Centro Interdisciplinario de Modelación Matemática y Computacional de la Universidad de Varsovia.

7) Arquitectura de extracción automática de metadatos basada en plantillas (Flynn, Zhou, Maly, Zeil, & Zubair, 2007): aborda directamente el problema de hacer frente a grandes colecciones heterogéneas de documentos con diversos diseños y

arquitecturas, dividiéndolo en partes más manejables. Propuesta en el Departamento de Ciencias de la Computación de la Universidad Old Dominion de la ciudad de Norfolk, Estados Unidos de América.

3.1 SAXEF: System for Automatic eXtraction of E-learning object Features

Este sistema permite extraer indicadores didácticos de páginas Web en forma automática, considerando cada página Web o grupo de páginas Web como un OA. Dichos indicadores ayudarán a los profesores a tomar la decisión de si el contenido de una página Web es útil para incluirla dentro de un curso en línea de determinada temática.

A partir de un curso o una única página Web, SAXEF genera una tarjeta de identificación de E-learning (EIC por sus siglas en inglés) que contiene la siguiente información del OA: Temas principales, temas secundarios, si el OA corresponde a contenido teórico o práctico, si es sintético o analístico, tipos y niveles multimedia, nivel de complejidad, enlaces a otros EIC con el mismo tema y enlaces a otro EIC con temas relacionados. La Tabla 3.1 muestra la información obtenida y la relación con los metadatos de DCMI y LOM.

Cabe resaltar que SAXEF no extrae metadatos bajo un estándar en particular, más sin embargo tal como se muestra en la Tabla 3.1, la mayoría de la información que contiene la EIC puede corresponderse con metadatos de los estándares LOM y/o DCMI. Adicionalmente, los metadatos obtenidos están casi todos directamente relacionados con el nivel educativo del OA, lo que hace que este sistema sea altamente utilizado en el contexto educativo, no solo por profesores que desean crear un nuevo curso en línea haciendo uso del material existente en Internet, sino también por estudiantes interesados en organizar de forma didáctica sus páginas Web de consulta.

La arquitectura de SAXEF está compuesta de tres niveles, bajo, medio y alto, tal como se muestra en la Figura 3.1:

- *Nivel bajo*: identifica y separa todos los componentes del OA, como por ejemplo, texto, imágenes, video, audio, entre otros.
- *Nivel medio*: extrae características específicas de cada componente, haciendo uso de algoritmos especializados.

- *Nivel alto*: fusiona las características extraídas en el nivel medio mediante la creación de las tarjetas de identificación EICs, que contienen la estructura e indicadores de los OAs. Esto implica también un análisis más profundo del contexto completo del OA.

Tabla 3.1. Metadatos contenidos en la EIC.

EIC	Descripción	DCMI	LOM
Temas principales	Expresa las claves o frases que describen el título o el contenido del recurso.	DC.Subject	1.5:General.Keyword o 9:Classification con 9.1:Classification.Purpose igual a "disciplina" o "idea".
Temas secundarios	Temas secundarios del contenido del recurso.	N/A	N/A
Teórico o práctico	La categoría del recurso.	DC.Type	5.2:Educational.Learning Resource Type
Sintético o analítico	Medida subjetiva de la utilidad educativa del recurso según su tamaño/duración.	N/A	5.4:Educational.Semantic Density
Tipos y nivel multimedia	En qué forma el usuario interactúa con el recurso (activo, expositivo, mixto).	N/A	5.1:Educational.Interactivity Type
Nivel de complejidad	Nivel de dificultad del recurso.	N/A	5.8:Educational.Difficulty
Enlaces a otros EIC con los mismos temas	Un identificador de un segundo recurso y su relación con el recurso actual. Este elemento permite enlazar los recursos relacionados y las descripciones de los recursos.	DC.Relation	7.2.2:Relation.Resource.Description
Enlaces a otros EIC con temas relacionados	Un identificador de un segundo recurso y su relación con el recurso actual. Este elemento permite enlazar los recursos relacionados y las descripciones de los recursos.	DC.Relation	7.2.2:Relation.Resource.Description

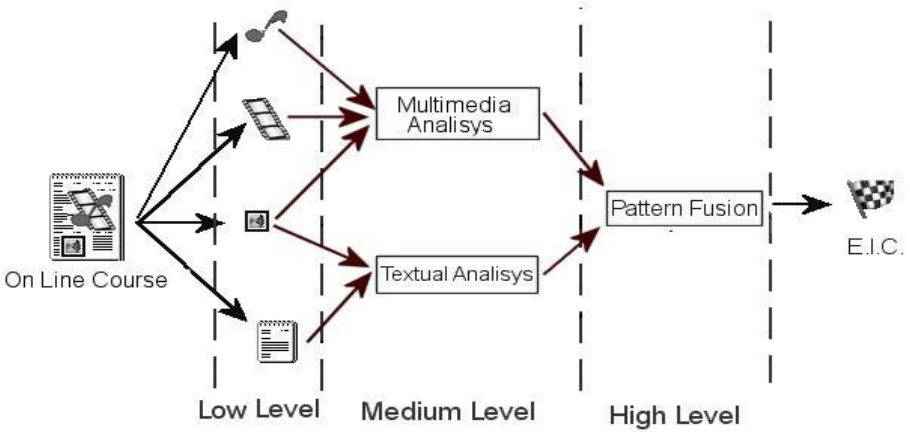

Fig. 3.1. Arquitectura SAXEF. Tomado de Figura 1. SAXEF architecture, (Alfano, Lenzitti, & Visalli, 2007)

Las técnicas y recursos utilizados en cada uno de los niveles se describen a continuación:

- *Nivel bajo*: el primer paso consiste en que el usuario ingrese la URL de la página Web o curso en línea que desea analizar; en ese momento, SAXEF toma el OA y analiza su código, que puede estar escrito en html, xhtml, asp, php, entre otros; a partir de este análisis, identifica y almacena en base de datos cada uno de los diferentes componentes de la página junto con la dirección de dicha página. Si el usuario está analizando un curso en línea completo, se analizarán entonces todas y cada una de las páginas que están referenciadas desde la página principal y que tengan la misma URL raíz.

- *Nivel medio*: en este nivel se realiza análisis de texto y análisis multimedia.

  Analizador de texto: se lleva a cabo sobre el texto de la página Web ejecutando los siguientes pasos:

  - Se eliminan todas las palabras comunes (artículos, preposiciones, pronombres, verbos comunes, etc.). Con este objetivo, se ha creado un archivo de texto que contiene la lista de esas palabras y que puede modificarse fácilmente a través de la interfaz web principal.

  - Se calculan las ocurrencias de palabras simples y las ocurrencias de parejas de palabras.

  - Se identifican las palabras dentro de las etiquetas "relevantes" <title> y <meta>.

  - Se encuentran las palabras más relevantes dentro del texto. Esto se logra mediante la poda del conjunto de palabras seleccionadas hasta ahora y considerando porcentajes específicos para las ocurrencias de palabras individuales y parejas.

  - Se asigna un peso a cada palabra seleccionada. En la práctica, el peso es un puntaje que la palabra obtiene dependiendo de cuándo y dónde aparece dentro del texto.

  Analizador multimedia: para cada página Web se calcula el tamaño de las áreas de texto y multimedia. El tamaño del área de texto se determina multiplicando el número de caracteres de la página Web por el área ocupada por cada carácter; se estima que

cada carácter de tamaño medio ocupa un área de alrededor de 100 píxeles. Se eligió un tamaño promedio para cada carácter porque la información sobre el tamaño de los caracteres no siempre está presente en el código de la página (por ejemplo, tal información puede estar contenida en hojas de estilo externas). El tamaño del área multimedia se determina sumando las áreas de los objetos multimedia presentes en la página Web. En particular, se consideran los tamaños (en píxeles) de imágenes, vídeos y animaciones. Por otra parte, si se encuentra un archivo de audio, se estima su tamaño (en bits) y dicho valor se divide por 16 bits (tamaño de muestreo).

Se han llevado a cabo un conjunto de experimentos en los dos tipos de analizadores para verificar su facilidad de uso y precisión de los resultados. Se ha visto que el analizador de texto es muy eficiente y proporciona resultados similares a los obtenidos a través del análisis humano. Por otro lado, los resultados del analizador multimedia son bastante completos pero pueden sintetizarse con más dificultad. Esto ocurre debido a la presencia potencial de elementos multimedia no relacionados con el contenido de la página (por ejemplo, banners de publicidad) que pueden llegar a alterar los resultados del análisis automático. Para superar este tipo de inconvenientes, ambos analizadores proporcionan al usuario la posibilidad de eliminar algunos de los resultados (en términos de palabras o elementos multimedia) y luego recalcular los indicadores EIC.

- *Nivel alto*: en este nivel, SAXEF calcula los indicadores que serán almacenados en la EIC de la siguiente manera:

  Temas principales y secundarios: a partir de los resultados del análisis de texto, se consideran temas principales las palabras con las dos puntuaciones más altas y temas secundarios las palabras con las siguientes cuatro puntuaciones más altas.

  Sintético o analítico: suponiendo que el área de la página Web es la suma del área de texto y del área multimedia, la relación entre el área de texto y el área total proporcionará el índice analítico (expresado como porcentaje); al mismo tiempo, la relación entre el área multimedia y el área total proporcionará el índice sintético. Aquel índice con mayor porcentaje determinará la densidad semántica.

Tipos y nivel multimedia: el nivel de multimedia (expresado como porcentaje) indica la presencia de los diferentes tipos de medios en una página Web (o curso). El índice de multimedia se calcula de la siguiente manera:

- Si sólo hay texto, el índice es igual al 20%.
- Si existen imágenes, el índice se incrementará de un porcentaje entre 0 y 20% proporcional al área de la imagen. Para áreas mayores de 20000 píxeles, el porcentaje permanecerá igual al 20%.
- Si existen audios, el índice se incrementará de 20% sólo si no hay archivos de vídeo.
- Si hay videos, el índice se incrementará en un 40%
- Si hay flash u otros tipos de animaciones, el índice aumentará 20%.
- El índice será igual al 100% cuando todos los tipos de medios multimedia estén presentes.

Aunque SAXEF está dando resultados bastante estables y confiables, es necesario llevar a cabo más pruebas para refinar los analizadores de texto y multimedia basados en los resultados. Además, es importante y necesario incluir en la EIC otros indicadores didácticos de interés para el usuario, lo que implica diseñar los analizadores correspondientes para extraerlos. Finalmente, se busca construir un motor de búsqueda de e-learning alrededor de SAXEF, con el fin de darle la posibilidad al usuario hacer peticiones en términos de indicadores didácticos y buscar automáticamente en Internet para encontrar las páginas web que mejor se adapten a los requisitos del usuario.

3.2 TWYS: Tang Way Yuen System

Esta propuesta se centra en la extracción automática de metadatos LOM de páginas Web HTML, teniendo en cuenta que estos OAs son adecuados para trabajar en un proyecto de investigación por la gran cantidad de recursos que se encuentra en Internet. Se considera una página Web HTML a puro contenido HTML, excluyendo información de otros OAs embebidos en la página tales como Word, Power Point, flash, entre otros.

Se eligió para trabajar el estándar de metadatos LOM, dado que éste fue desarrollado específicamente para el dominio de la educación y los cerca de 80 metadatos que plantea

conducen a una buena descripción de los OAs, que en consecuencia ayuda a los usuarios a identificar fácilmente recursos relevantes para un propósito determinado. Para tal fin, se clasificaron y reorganizaron los elementos LOM en 4 grupos, que son los que se muestran en la Tabla 3.2

No obstante, como trabajo futuro los autores proponen automatizar la extracción de otro estándar de metadatos tal como Dublin Core haciendo la traducción de LOM a Dublin Core o haciendo un estudio de la clasificación de Dublin Core para el proceso de extracción automática.

Tal como se muestra en la Tabla 3.2, algunos elementos LOM requieren ser establecidos por los autores o usuarios mientras que parte de la información encontrada en una página Web HTML puede ser directamente mapeada con elementos LOM; para esto último se requieren métodos de mapeo directo y reglas heurísticas con el fin de determinar el valor de los metadatos.

- *Mapeo directo*: se refiere a que la información HTML puede ser mapeada directamente con metadatos bajo el estándar LOM. Para esto se adoptaron algunas técnicas bien conocidas como son Stop Words, Term Frequency Weigthing (TF-IDF) y Ontología.

  Stop Words y TF-IDF son técnicas de recuperación de información mediante la eliminación de información poco útil, con el fin de centrarse en la información importante del documento. Ontología hace referencia a la especificación explícita de una conceptualización ayudando a encontrar la relación de una página Web HTML con alguna ruta de concepto predefinida.

- *Reglas heurísticas*: algunos metadatos LOM no pueden ser obtenidos directamente de la información encontrada en las páginas Web HTML, particularmente los de tipo educacional en el grupo relacionado con el contenido. Sin embargo, esos valores pueden ser deducidos de los datos HTML a través de algunas reglas de conteo o comparación. De aquí se derivan dos reglas heurísticas para determinar dichos valores por mapeo:

  - "Chequear la existencia de tag HTML": se refiere a chequear si cierto tag HTML existe en la página Web.

- "Conteo estadístico de HTML y Contenido": se refiere al conteo de la cantidad de ciertos tags HTML y palabras en una página Web.

Tabla 3.2. Agrupación de elementos LOM para TWYS.

Grupo	Descripción	Elementos LOM	Método de Extracción
Control de Usuarios/Autores	Información de la instancia del metadato en sí misma. Derechos de propiedad intelectual y condiciones de uso del OA. Relaciones entre OA. Información del uso educacional del OA y acerca de cuándo y por quienes fueron creados los comentarios.	Meta-Metadata (LOM 3) Rigths (LOM 6) Relation (LOM 7) Annotation (LOM 8)	Todos los elementos LOM de esta categoría requieren ser completados por el autor o usuario y no pueden ser extraídos automáticamente
Relacionado con Propiedades Físicas	Información general que describe el OA como un todo. Características relacionadas con la historia y estado actual del OA. Requerimientos y características técnicas del OA.	General (LOM 1) Life Cycle (LOM 2) Technical (LOM 4)	Muchos de los elementos LOM de este grupo pueden ser extraídos fácilmente del OA, a través del contenido HTML, tags HTML o encabezados HTTP. Algunos otros no pueden ser extraídos automáticamente y requieren ser completados por el autor o usuario
Relacionado con el contenido	Relacionados con el contenido del OA. Características educacionales y pedagógicas del OA. Descripción del OA en relación a un sistema de clasificación particular.	General (LOM 1) Educational (LOM 5) Classification (LOM 9)	Muchos de los elementos LOM de este grupo pueden ser extraídos fácilmente del OA, a través de mapeo o métodos de extracción y ontologías. Algunos requieren ser completados por el autor o usuario.
Relacionado con el uso	Requerimientos y características técnicas del OA. Uso educacional del OA.	Technical (LOM 4) Educational (LOM 5)	Todos los elementos LOM de esta categoría requieren ser llenados por el autor o usuario

Teniendo en cuenta lo anterior, se diseñaron 3 reglas de mapeo directo (extracción directa, uso de TF-IDF, uso de ontología) y 4 reglas heurísticas (chequeo de la existencia de tags HTML, conteo estadístico de tags HTML, conteo estadístico del total de palabras y multimedia, y conteo estadístico de palabras distintas), para determinar los valores de los elementos LOM, tal como muestra la Tabla 3.3.

Tabla 3.3. Resumen de reglas de mapeo de elementos LOM con datos HTML.

Reglas de Mapeo	Método de Mapeo	Dato HTML para referencia de mapeo	Elemento(s) LOM extraídos/deducidos
Mapeo directo	Extracción directa	URL	Entry (LOM 1.1.2)
			Location (LOM 4.3)
		Tag HTML <Title>	Title (LOM 1.2)
		Encabezado HTTP "Content-Language"	Language (LOM 1.3)
		Tag HTML <a href=mailto>	Entity (LOM 2.3.2)
		Encabezado HTTP "Last-Modified"	Date (LOM 2.3.3)
		Encabezado HTTP "Content-Type"	Format (LOM 4.1)
		Encabezado HTTP "Content-Length"	Size (LOM 4.2)
Mapeo directo	Stop Word, TF-IDF	Contenido de la página web HTML	Description (LOM 1.4)
			Keyword (LOM 1.5)
Mapeo directo	Ontología	Contenido de la página web HTML	Purpose (LOM 9.1)
			ID (LOM 9.2.2.1)
			Entry (LOM 9.2.2.2)
			Description (LOM 9.3)
			Keyword (LOM 9.4)
Regla heurística 1	Chequeo de la existencia de tags HTML	Tag HTML <Form action=>, <Input>	Interactivity Type (LOM 5.1)
Regla heurística 2	Conteo estadístico de tags HTML	Tag HTML <Input>, <action=>	Interactivity Level (LOM 5.3)
Regla heurística 3	Conteo estadístico del total de palabras y multimedia	Conteo total de palabras de la página web HTML, Conteo total de multimedia de la página web HTML	Semantic Density (LOM 5.4)
Regla heurística 4	Conteo estadístico de palabras distintas	Conteo de las palabras distintas del contenido de la página web HTML	Difficulty (LOM 5.8)

El framework de extracción automática TWYS es ilustrado en la Figura 3.2, donde se distinguen claramente 5 módulos principales: Crawler, Stop Words, Ontología, Parser HTML y Extracción LOM.

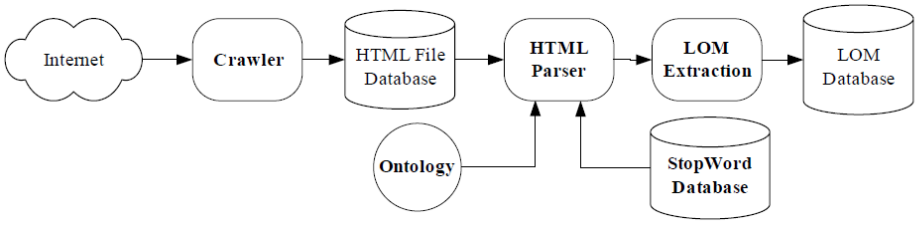

Fig. 3.2. Framework de Extracción Automática de Elementos LOM.

El proceso de extracción automática de elementos LOM de páginas Web HTML se resume de la siguiente manera:

1. Coleccionar automáticamente páginas Web HTML de Internet relacionadas con el dominio de interés, haciendo uso del Crawler, y almacenarlas en la BD local de archivos HTML.

2. Ejecutar el Parser HTML para procesar las páginas web HTML coleccionadas y separar el contenido HTML de sub-enlaces, encabezados y tags HTML.

3. Eliminar información poco útil del contenido HTML mediante el uso de Stop Words.

4. Determinar cuáles de las páginas coleccionadas están relacionadas con el dominio de interés, mediante el uso de la ontología.

5. Extraer y generar registros LOM correspondientes al contenido HTML válido haciendo uso de las reglas de mapeo y métodos diseñados en el módulo de Extracción LOM.

6. Almacenar los registros LOM extraídos en la BD LOM para uso posterior.

El diseño del framework de extracción automática de LOM está basado en el paradigma orientado a objetos para mayor facilidad en el mantenimiento y extensión. El framework está construido por 3 capas, cada una enfocada en una función específica, incluyendo la capa de preparación y planeación requerida para la extracción de LOM, la capa de extracción para la extracción de registros LOM del contenido HTML obtenido de Internet por el sistema, y la capa de presentación para la revisión de OA's relevantes por los usuarios finales a través de los registros LOM en el sistema.

TWYS fue ejecutado con éxito para la extracción automática de metadatos LOM de cerca de 3000 página Web HTML obtenidas de Internet, relacionadas con el dominio de las matemáticas. El framework de extracción automática puede generar correctamente los registros LOM correspondientes siempre y cuando la información necesaria esté presente en los encabezados HTTP, etiquetas HTML y contenido de la página Web. Las pruebas de concepto realizadas con 29 profesores de secundaria para validar la exactitud de los metadatos LOM extraídos por el framework, indican que dichos metadatos son generalmente precisos y razonables.

3.3 LookIng4LO: Sistema Informático para la Extracción Automática de Objetos de Aprendizaje

A diferencia de las dos propuestas presentadas anteriormente, LooKIng4LO recibe como entrada documentos no estructurados como por ejemplo archivos PDF, TXT y paquetes SCORM, y extrae información a partir de un área temática y un conjunto de componentes pedagógicos, empaquetándola nuevamente en Objetos de Aprendizaje (OAs) donde cada uno de ellos posee metadatos que lo describen. El área temática se define a través de una ontología y los componentes pedagógicos son modelados con reglas para definir patrones de búsqueda. Los OAS generados son empaquetados utilizando el estándar SCORM.

Lo más destacado de esta propuesta es la consideración del diseño pedagógico para la generación de metadatos adecuados. La Figura 3.3 representa los principales componentes del proceso de generación de OAs con metadatos:

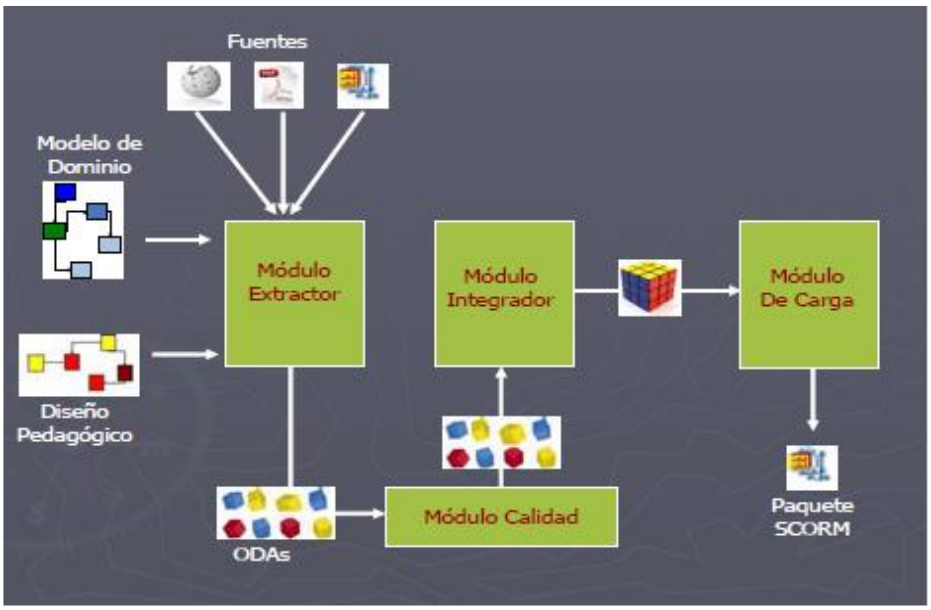

Fig. 3.3. Arquitectura General de LooKIng4LO.

- *Documento o Fuente*: cualquier elemento digital a partir del cual se puedan generar OA's. En principio, LooKIng4LO está diseñado para trabajar con documentos no estructurados como por ejemplo página Web HTML, archivos PDF, texto y paquetes SCORM; sin embargo, dada la gran variedad de fuentes posibles, se diseñó el sistema de forma que pueda evolucionar a nuevos formatos y estrategias de extracción.

- *Modelo de Dominio*: su función consiste en definir cualquier objeto o entidad que se quiera representar, y se utiliza para modelar el tema sobre el que se busca generar OA's. Define qué tema se quiere buscar y se modela a través de una ontología liviana.

- *Diseño Pedagógico*: indica las características pedagógicas buscadas sobre la fuente y se modela mediante un conjunto de reglas.

- *Objeto de Aprendizaje (OA)*: es modelado como un elemento que contiene texto más una estructura (árbol n-ario) de metadatos que lo describe. Cada elemento de esta estructura de metadatos, tiene un nombre, valor y un conjunto elementos hijo del mismo tipo. Esta estructura permite manejar metadatos definidos en formato LOM y extensiones realizadas sobre esta.

- *Módulo Extractor*: recibe dos entradas, el *modelo de dominio* y el *diseño pedagógico*, que determinan si la información de cierta fuente clasifica o no dentro de las características pedagógicas buscadas. Los OA's resultantes serán aquellos que cumplan con las condiciones planteadas por estas dos entradas y que estarán enriquecidos con metadatos extraídos automáticamente que pasan a ser analizados por el *módulo de calidad*.

- *Módulo de Calidad*: se encarga de medir la calidad de los OAs en función de un conjunto de factores de calidad que son determinados por quien hace uso del sistema, como por ejemplo, fecha de última actualización o el grado de accesibilidad del material. A la salida del procesamiento del módulo, los OAs contarán con nuevos metadatos sobre la calidad de los mismos.

- *Módulo Integrador*: en este módulo el usuario podrá revisar los ODAs extraídos y anotados con metadatos de calidad (salida del *módulo de calidad*), y en función de esto, seleccionar cuáles han de ser persistidos por el módulo de carga y cuáles no.

- *Módulo de Carga*: los OAs seleccionados en el *módulo integrador* junto a los metadatos obtenidos, tanto de forma automática como manual, son tomados por el módulo de carga y empaquetados utilizando el estándar SCORM.

En LooKIng4LO un OA tiene cuatro conjuntos de metadatos que clasifican la información que lo describe de acuerdo al origen de donde es obtenida (Tabla 3.4).

El prototipo inicial de LooKIng4LO extrae un tipo de metadato específico para cada uno de los componentes pedagógicos que trabaja y un metadato general que aplica a todos (Tabla 3.5). La elección de estos metadatos fue arbitraria, ya que se buscó mostrar la utilidad de la clasificación de los metadatos específicos y generales, e implementar su extracción para

demostrar su factibilidad. Extender y/o incorporar nuevos metadatos específicos/generales implica implementar nuevas reglas.

Tabla 3.4. Clasificación de los metadatos extraídos por LooKIng4LO.

Clasificación	Descripción
Fuente	Metadatos disponibles a nivel de cada fuente o recurso. Se refiere a la información asociada al archivo, como autor, fecha de creación, etc.
Generales	Son generados automáticamente por el sistema y contienen información sobre el contenido del documento, como el idioma.
Específicos	Generados automáticamente y son específicos a un tipo de componente pedagógico (ejercicios, ejemplos, definiciones, etc.).
Externos	Se añaden en forma manual por el usuario del sistema. Se asocian a todos los OA's generados durante la ejecución, y para estos, se debe proporcionar su nombre y valor.

Tabla 3.5. Componentes pedagógicos y metadatos específicos/generales extraídos.

Componente Pedagógico	Metadato específico	Metadato general
Definición	*Tiempo de lectura*: estimación del tiempo que requiere leer el contenido del OA. Se calcula contabilizando la cantidad de palabras de contenido del OA, dividido por una constante (que para este sistema es 200).	*Autor* Se busca a nivel de todo el documento y no solo en el contenido de un tipo de componente pedagógico en particular. Cuando se identifica el o los autores de un documento, se extrae también el correo electrónico y página web de cada autor en caso de que esta información esté disponible.
Ejemplo	*Tiene imagen*: asocia un valor booleano, que es verdadero en caso de que el OA contenga una imagen o figura como parte de su contenido, o falso en caso contrario	
Ejercicio	Nivel de interactividad: asocia un valor entero al OA, dependiendo de si el ejercicio debe enviarse por email (asigna 5), a un foro (asigna 9), news (asigna 8) o no se requiere ninguna de estas actividades para su resolución (asigna 0).	

A continuación se detalla un poco más el módulo extractor, con el fin de conocer las técnicas y recursos utilizados en el proceso de extracción. El módulo extractor trabaja en dos niveles, el primero donde se definen, a través de una ontología, los componentes pedagógicos que se quieren extraer, como por ejemplo, definición, ejercicio, ejemplo, entre otros; el segundo nivel se encarga de identificar los metadatos asociados a cada uno de los componentes definidos en el primer nivel y que deben estar incluidos en el diseño pedagógico que se va a utilizar.

Se trabaja con las características del diseño pedagógico escrito en reglas de negocio usando el motor de reglas Drools (JBoss, Drools, 2006). Drools (conocido también como JBoss Rules) es

un motor de Reglas de Negocio implementado totalmente en Java, gratuito y encadenado hacia delante. Los objetos que se crean se van almacenando en una memoria de trabajo, y estos objetos son luego evaluados para saber qué reglas cumplen y cuáles no. Aquellas reglas que se cumplan harán que el módulo ejecute las instrucciones requeridas para la extracción de información, poniendo especial atención desde el punto de vista pedagógico la categoría Educational de LOM que traen los OA's empaquetados en el Standard SCORM.

Dentro de la categoría Educational se encuentra el metadato Learning Resource Type, donde se distinguen los siguientes tipos:

- *Gráfico*: diagram, graph, slide, figure.
- *Texto*: exercise, simulation, questionnaire, index, table, narrative text, exam, experiment, problem statement, self assessment, lecture, definition, example, FAQ, theorem, activity, conclusion, demonstration, objective, midterm examination.

Adicionalmente, se agregan los tipos que son significativos para la organización del curso:

- *Cartelera*: notice board.
- *Horarios*: time-table.
- *Programa*: program.
- *Cronograma*: scheduler.
- *Entrega de tareas online*: automatic receiver.

Para la identificación de mails, foros y newsgroup usa VCARD, que es un formato estándar para el intercambio de información personal.

Teniendo en cuenta lo anterior, LooKIng4LO es altamente aplicable en contextos educativos, dado que uno de los componentes principales es el diseño pedagógico; adicionalmente, se diseñó pensando en extender su uso hacia otros tipos de formatos de fuentes. El repositorio de pruebas de este sistema se enfocó en materiales de cursos abiertos depositados en un sitio OCW (Open Course Ware por sus siglas en inglés); allí se encuentra una buena fuente de datos de prueba para la herramienta como son: planificación de cursos (programas, temarios, objetivos pedagógicos, calendarios, etc.), contenidos (bibliografía, documentos, material

audiovisual, material auxiliar, etc.) y distintas actividades pedagógicas (ejercicios, tests, proyectos, prácticas de laboratorio, etc.).

3.4 MAGIC: Metadata Automated Generation for Instructional Content

Como su nombre lo indica, esta propuesta genera automáticamente metadatos para contenido instructivo, conforme al estándar SCORM. Dentro del contenido instructivo se contemplan página Web HTML, tipos de archivos no estructurados tales como TXT, PDF y PPT y archivos de tipo video y audio (AVI, MPG, MP3, MP4, WMA).

Con MAGIC, los desarrolladores de contenido instructivo pueden generar y editar metadatos SCORM, y describir ricamente su contenido para habilitar su uso en aplicaciones de aprendizaje distribuido. En la Figura 3.4 se muestra la arquitectura del sistema MAGIC que consiste en un Ambiente de Generación Metadatos (MGE) y un Editor de Metadatos. El proceso general de uso del sistema es el siguiente:

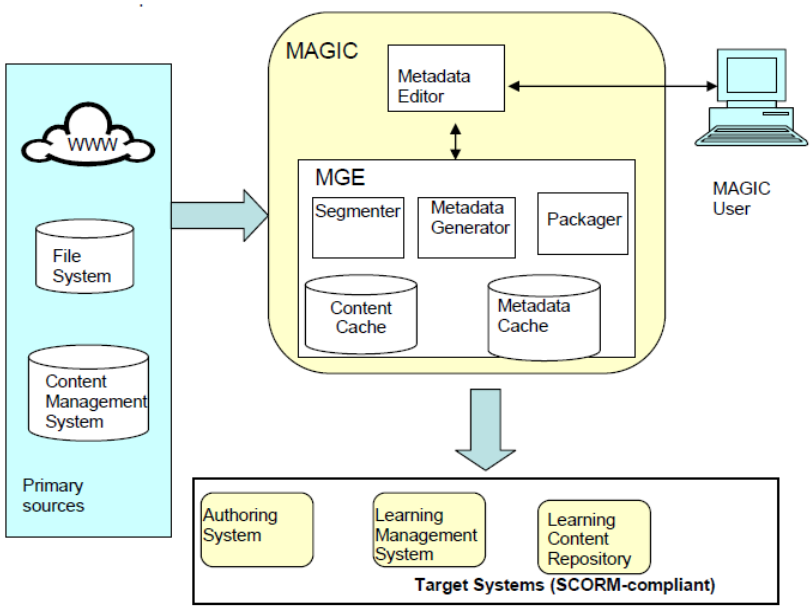

Fig. 3.4. Arquitectura del Sistema MAGIC. Tomado de Figura 1 (Li, Dorai, & Farrell, 2005)

1. El usuario MAGIC (un autor o un desarrollador de cursos) interactúa con el sistema mediante el Editor de Metadatos. El usuario accede a un documento de formación o un video instructivo introduciendo una dirección URL en la Web, el nombre de un archivo local o una dirección en un Sistema de Gestión de Contenidos.

2. El Editor de Metadatos carga el contenido en Caché de Contenido donde es opcionalmente revisado por el Segmentador. El usuario puede entonces ver el recurso resultante en el Editor de Metadatos para sugerir puntos de segmentación.

3. El recurso luego es procesado por el Generador de Metadatos, lo que crea un registro de metadatos SCORM y lo almacena en el Caché de Metadatos. Posteriormente, el usuario puede ver y corregir el registro SCORM.

4. Finalmente, el usuario puede solicitar al Empaquetador crear un paquete compatible con SCORM (con un Manifiesto IMS) que puede exportarse a sistemas de creación SCORM, sistemas de gestión de aprendizaje, o repositorios de aprendizaje de contenido.

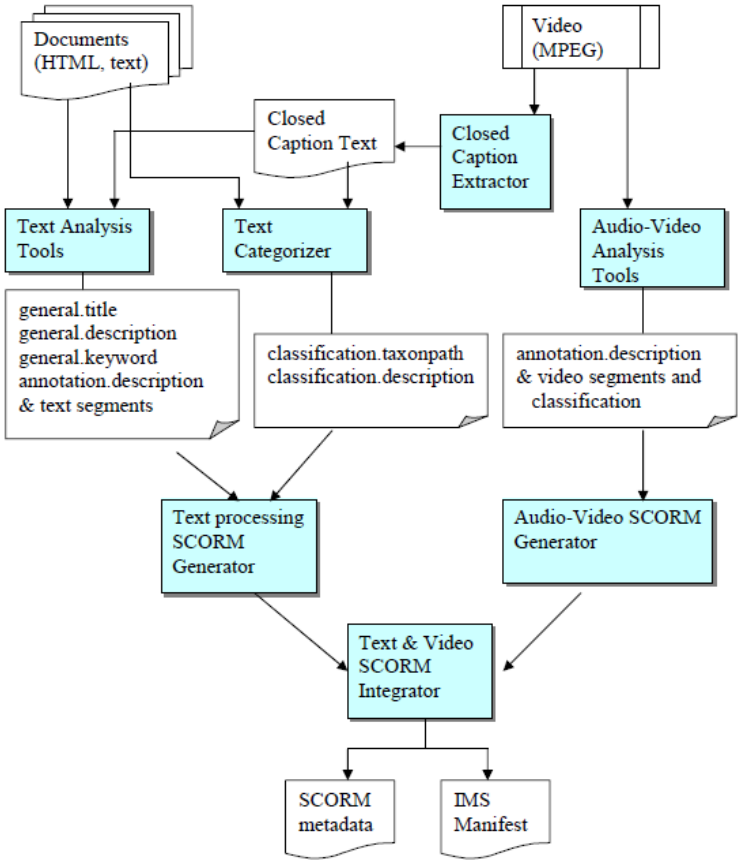

Fig. 3.5. Herramientas de Procesamiento de Texto y Video. Tomado de Figura 2 (Li, Dorai, & Farrell, 2005)

A continuación se detalla un poco más el componente generador de metadatos, con el fin de conocer las técnicas y recursos utilizados en el proceso de extracción. El generador de metadatos consiste en el grupo de herramientas de procesamiento de texto y video que se

muestran en la Figura 3.5, y que se han desarrollado e integrado a través de un conjunto común de interfaces de programas de aplicación:

- *Herramientas de análisis de texto*: están diseñadas para explorar a través de documentos HTML o PDF de gran tamaño e identificar o extraer información útil a través del siguiente procedimiento:

  - Primero tokenizan (dividen) el documento usando Frost (IBM, IBM LanguageWare Resource Workbench, s.f.).

  - Luego, cada palabra tiene una etiqueta asignada con su parte en la oración (por ejemplo, nombre, verbo o adjetivo).

  - A continuación, un conjunto de módulos de análisis se aplican para extraer la información que se muestra en la Tabla 3.6 y que corresponde a los metadatos que luego serán utilizados por el generador SCORM.

Tabla 3.6. Metadatos extraídos del análisis de texto

Descripción	DCMI	LOM
El nombre dado a un recurso, normalmente proporcionado por el autor.	DC.Title	1.2:General.Title
Palabras clave técnicas que se clasifican de la más específica a la más genérica	DC.Subject	1.5:General.Keyword o 9:Classification con 9.1:Classification.Purpose igual a "disciplina" o "idea".
Palabras clave de entidad, que incluyen personas, lugares y nombres de organizaciones, todas calificadas por la frecuencia	N/A	9.4:Clasiffication.Keyword
Límites de cambio de tema	N/A	N/A
Descripción resumida que comprende un par de frases importantes	DC.Description	1.4:General.Description

  - Por último, el componente generador SCORM genera un archivo de metadatos LOM poblando elementos de metadatos adecuados en formato XML utilizando la información extraída anteriormente.

- *Herramientas de categorización de texto*: incluye una taxonomía de alta cobertura, independiente del dominio, y un componente clasificador de texto que de forma automática y precisa asigna documentos de texto a determinadas categorías de esta taxonomía. Esto se hace comparando el documento con los modelos que han sido

previamente diseñados para un gran número de temas, también llamados centroides, que son modelos de archivos precalculados de la forma de vectores numéricos que describen la palabra que se deriva de un documento típico en su categoría.

- *Herramientas de análisis de audio y video*: teniendo en cuenta que el procesamiento de archivos de audio y video es mucho más complejo, estas herramientas segmentan el vídeo en unidades pequeñas, identifican su semántica de contenido mediante la detección de un narrador, reconociendo tipos de sonido y extrayendo texto del vídeo, y finalmente los etiquetan con anotaciones que describen los elementos narrativos en los segmentos.

En general, esta propuesta presenta un sistema que puede identificar automáticamente los segmentos y generar metadatos SCORM críticos para el contenido educacional, lo que ofrece un marco y una solución que se puede utilizar por una comunidad educativa de usuarios, que incluye fácil reutilización de contenido, mejora de la interoperabilidad, y el registro más puntual de los contenidos para los desarrolladores de cursos y autores de contenido. Para lograr estos objetivos, se han desarrollado diversas herramientas interpretativas de texto, audio y análisis visual, lo que implica grandes esfuerzos de diseño y procesamiento al tratarte de tipos de archivos cuyos metadatos no son comúnmente extraídos.

3.5 Asistente para el Depósito de Objetos en Repositorios con Extracción Automática de Metadatos

En este trabajo 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, con el fin de incorporarlo a RepHip el Repositorio Hipermedial institucional de la Universidad Nacional de Rosario, Argentina.

La arquitectura propuesta para este proceso se muestra en la Figura 3.6 y los pasos correspondientes se describen a continuación:

1. *Submit Step*: se refiere al paso donde el usuario elige la colección en la cual quiere depositar el documento, acepta la licencia institucional y carga el(los) archivo(s) asociado(s) al OA, cuyo documento principal se envía al asistente para la extracción de metadatos.
2. *PDF/Word to TXT*: al archivo enviado en formato PDF o Word se le extrae el contenido en formato txt.

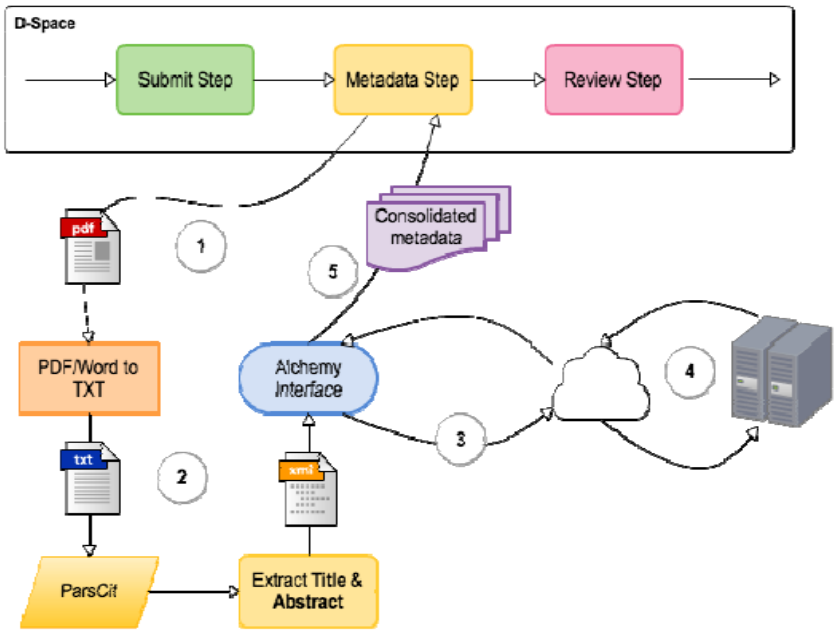

Fig. 3.6. Arquitectura propuesta para el proceso de extracción semiautomática. Tomado de Figura 2 (Casali, Deco, Bender, Fontanarrosa, & Sabater, 2013)

3. *ParsCit- Extract Title & Abstract*: el texto se pasa al analizador de estructura, que procesa el contenido y genera un archivo XML donde se identifica título, resumen y el contenido separado por páginas.
4. *Alchemy Interface*: el archivo XML se envía a *Alchemy* (IBM, AlchemyAPI, 2009) para la extracción de metadatos. Alchemy envía la información a su servidor, extrayendo el idioma y las palabras claves.
5. *Consolidated metadata*: se consolida la respuesta del servidor y se envía a DSpace (DuraSpace, 2002) un archivo XML con todos los metadatos extraídos automáticamente.
6. *Metadata Step*: los metadatos extraídos automáticamente se presentan al usuario por medio de una interfaz, para que sea validados y completados en el proceso de descripción del objeto.
7. *Review Step*: se realiza una verificación de todos los datos cargados.

Dado que el asistente está diseñado para ser incorporado en un RI construido bajo la plataforma DSpace los metadatos extraídos se rigen bajo el estándar DCMI (ver Tabla 3.7); sin embargo, tal como se revisó en la sección 2.2.2, actualmente existe la posibilidad de hacer equivalencia entre los estándares IEEE LOM y DCMI, lo que se considera una ventaja a la hora de expandir la usabilidad de esta propuesta.

La novedad de esta propuesta radica principalmente en el reordenamiento y modificación de los pasos del depósito de OA en RI, donde se distingue claramente la descripción de metadatos obligatorios y opcionales de acuerdo a la colección elegida para almacenar el OA y apoyado por el asistente de extracción automática; así mismo, en el último paso se da la posibilidad al usuario de modificar la información, conforme a su criterio. Uno de los puntos de mejora de esta propuesta consiste en poder procesar los archivos que presentan un formato que no permita extraer y transformar el contenido del mismo a texto plano, ya que dichos archivos no pueden ser tomados en cuenta en el proceso actual.

Tabla 3.7. Metadatos extraídos por el Asistente.

Descripción	DCMI	LOM
El nombre dado a un recurso, normalmente proporcionado por el autor.	DC.Title	1.2:General.Title
La persona u organización responsable de la creación del contenido intelectual del recurso.	DC.Creator	2.3.2:Life Cycle.Contribute.Entity cuando 2.3.1:Life Cycle.Contribute.Role tiene como valor "autor".
Expresa las claves o frases que describen el título o el contenido del recurso.	DC.Subject	1.5:General.Keyword o 9:Classification con 9.1:Classification.Purpose igual a "disciplina" o "idea".
Una descripción textual del recurso, como puede ser un resumen o una descripción de su contenido.	DC.Description	1.4:General.Description
Lengua/s del contenido intelectual del recurso.	DC.Language	1.3:General.Language

Aunque este asistente fue construido exclusivamente para RepHip, su diseño constituyó una base firme para la elaboración del algoritmo AMELOIR que será presentado en el capítulo 4.

3.6 CERMINE: Content ExtRactor and MINEr

El flujo de trabajo para el proceso de extracción de metadatos y referencias de literatura científica de documentos de origen digital en formato PDF, consta de 3 partes principales tal como se muestra en la Figura 3.7:

1. *Estructura básica de extracción*: toma un archivo PDF como entrada y produce una estructura jerárquica geométrica que representa el documento. La estructura se compone de páginas, zonas, líneas, palabras y caracteres. Cada zona está etiquetada con una de las cuatro categorías generales: METADATOS, REFERENCIAS, CUERPO y OTROS.

2. *Extracción de metadatos:* analiza las zonas etiquetadas como METADATOS y extrae en un archivo XML un rico conjunto de metadatos del documento.

3. *Extracción de referencias*: analiza las zonas etiquetadas como REFERENCIA y el resultado es un archivo XML con una lista de referencias bibliográficas de documentos. Dado que esta propuesta está diseñada para el análisis de publicaciones científicas, la extracción de referencias bibliográficas es de vital importancia en pro de ser utilizado en bibliotecas digitales que hagan parte de una red científica de conocimiento.

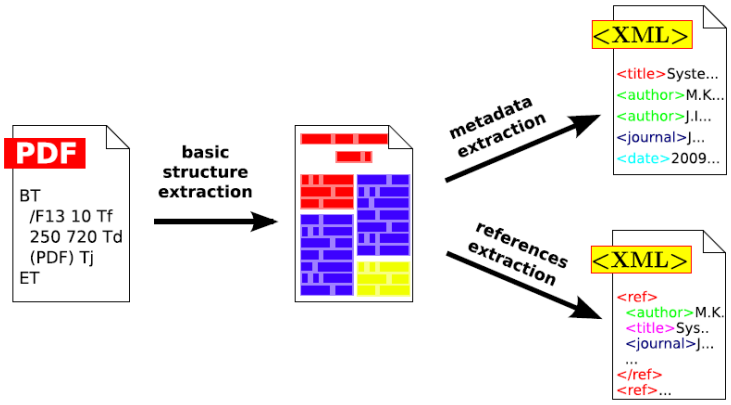

Fig. 3.7. Flujo de Trabajo CERMINE. Tomado de Figura 1 (Tkaczyk, Szostek, Dendek, Fedoryszak, & Bolikowski, CERMINE - Automatic extraction of metadata and references from scientific literature, 2014)

El resultado del flujo de trabajo CERMINE es un archivo JATS (NLM, s.f.) (Journal Article Tag Suite por sus siglas en inglés) que contiene un amplio conjunto de elementos y atributos XML para describir publicaciones científicas y es una aplicación de la norma NISO Z39.96-2012 (NISO, 2012). El listado de metadatos extraídos se muestra en la Tabla 3.8; a pesar de que la propuesta no sigue ningún estándar de metadatos en particular, la información obtenida puede ser fácilmente mapeada a DCMI y/o IEEE LOM.

En esta propuesta la implementación de la extracción de metadatos está basada en el uso de Máquinas de Soporte Vectorial (SVM por sus siglas en inglés) que son un conjunto de algoritmos de aprendizaje automático y supervisado, utilizados para resolver problemas de clasificación y regresión. Específicamente CERMINE hace uso de LIBSVM (Chang & Lin, 2011), un software integrado SVM cuya más reciente versión (versión 3.22) data de Diciembre 22 de 2016, para clasificar la zona general de METADATOS en clases de metadatos específicas, cada una de las cuales representa uno de los metadatos que se encuentra en la Tabla 3.8.

Tabla 3.8. Metadatos extraídos por CERMINE.

Campo	Descripción	DCMI	LOM
Título	El nombre dado a un recurso, normalmente proporcionado por el autor.	DC.Title	1.2:General.Title
Autor	La persona u organización responsable de la creación del contenido intelectual del recurso.	DC.Creator	2.3.2:Life Cycle.Contribute.Entity cuando 2.3.1:Life Cycle.Contribute.Role tiene como valor "autor".
Palabras Clave	Expresa las claves o frases que describen el título o el contenido del recurso.	DC.Subject	1.5:General.Keyword o 9:Classification con 9.1:Classification.Purpose igual a "disciplina" o "idea".
Resumen	Una descripción textual del recurso, como puede ser un resumen o una descripción de su contenido.	DC.Description	1.4:General.Description
Identificador del Recurso DOI	Secuencia de caracteres usados para identificar unívocamente un recurso. Por ejemplo la URL del recurso.	DC.Identifier	1.1.2:General.Identifier.Entry
Fecha	Una fecha en la que el recurso se puso a disposición del usuario en su forma actual.	DC.Date	2.3.3:Life Cycle.Contribute.Date cuando 2.3.1:Life Cycle.Contribute.Role tiene como valor "editor".
Editor	La entidad responsable de hacer que el recurso se encuentre disponible en la red en su formato actual.	DC.Publisher	2.3.2:Life Cycle.Contribute.Entity cuando 2.3.1:Life Cycle.Contribute.Role tiene como valor "editor".
Tipo de Recurso	La categoría del recurso.	DC.Type	5.2:Educational.Learning Resource Type

Posteriormente, se extrae información atómica de cada una de las zonas específicas etiquetadas haciendo uso de un conjunto simple de reglas heurísticas de mapeo, tales como, concatenación, diccionario de tipos, división usando lista de separadores, índices y distancias y expresiones regulares.

Esta propuesta es open source, fácil de adaptar a nuevos diseños de documentos y aplicable a cualquier contexto científico; su incorporación a entornos educativos como los RI estará sujeto a la viabilidad del uso de estándares de metadatos, ya que la evaluación frente a un conjunto de datos amplio y diverso muestra buenos resultados para los pasos individuales y para todo el flujo de extracción.

3.7 Arquitectura de Extracción Automática de Metadatos basada en Plantillas

La arquitectura de extracción automática de metadatos basada en plantillas aborda directamente el problema de hacer frente a grandes colecciones heterogéneas de documentos PDF, con diversos diseños y arquitecturas, dividiéndolo en dos partes más manejables:

1) Un nuevo documento es clasificado, asignándolo a un grupo con otros documentos de diseño similar. El objetivo es agrupar los documentos en los cuales metadatos como el autor y organización que lo publica, entre otros, parezcan similares.
2) Asociar a cada grupo de documentos una plantilla, una descripción que indique cómo agrupar bloques de texto en el diseño con bloques de metadatos. Por ejemplo, una plantilla puede indicar que el texto que figura con el tipo de letra más grande en la mitad superior de la primera página, es el título del documento.

El diseño de esta propuesta consta de un software de dominio comercial y público sumado a componentes desarrollados por el equipo de trabajo. La Figura 3.8 muestra el proceso completo, que se describe a continuación:

1) La entrada del proceso son documentos en formato PDF, con texto y/o imágenes escaneadas. Algunos documentos pueden contener una *Página de Reporte del Documento* (RDP Report Document Page por sus siglas en inglés), una de las varias formas estandarizadas que se inserta en el documento cuando se añade a la colección y que es clave a la hora de realizar el proceso de extracción de metadatos.
2) Los documentos ingresan al *procesamiento de entrada & OCR***,** donde se dividen en secciones procesados por un programa de Reconocimiento Óptico de Caracteres (OCR) y se convierten a formato estándar XML.
3) La primera etapa de extracción busca cualquier forma RDP presente en el documento; si el documento no contiene dicha página, ingresa al *proceso de extracción no-forma*, los demás ingresan al *proceso de extracción forma*.
4) El *proceso de extracción forma* se basa en el diseño regular que presenta una forma RDP; el procesador está poblado con una plantilla desarrollada para cada versión de forma RDP encontrada en la colección. Se ejecuta el proceso de extracción contra el documento utilizando cada una de las plantillas y se selecciona la plantilla que devuelve los mejores resultados. Si el procesador de forma no coincide con alguna plantilla el documento entra en el *proceso de extracción no-forma* descrito a continuación.

5) El *proceso de extracción no-forma* genera una solución de extracción candidata a las plantillas disponibles. El motor de extracción no-forma también usa plantillas de extracción basadas en reglas para localizar y extraer los metadatos. Cada plantilla contiene un conjunto de reglas diseñadas para extraer metadatos a partir de una sola clase de documentos similares.
6) Después del procesamiento y la extracción, los metadatos extraídos entran al *procesador de salida,* compuesto por un *módulo de post-procesamiento* y un *módulo de validación*.
7) El *módulo de post-procesamiento* se encarga de la limpieza y la normalización de los metadatos, conforme a la plantilla asociada al documento.
8) El paso final automatizado del proceso es el *módulo de validación* que, utilizando una matriz determinística y pruebas estadísticas, determina la aceptabilidad de los metadatos extraídos.
9) Cualquier documento que no cumple con los criterios de validación es marcado para revisión y corrección humana.

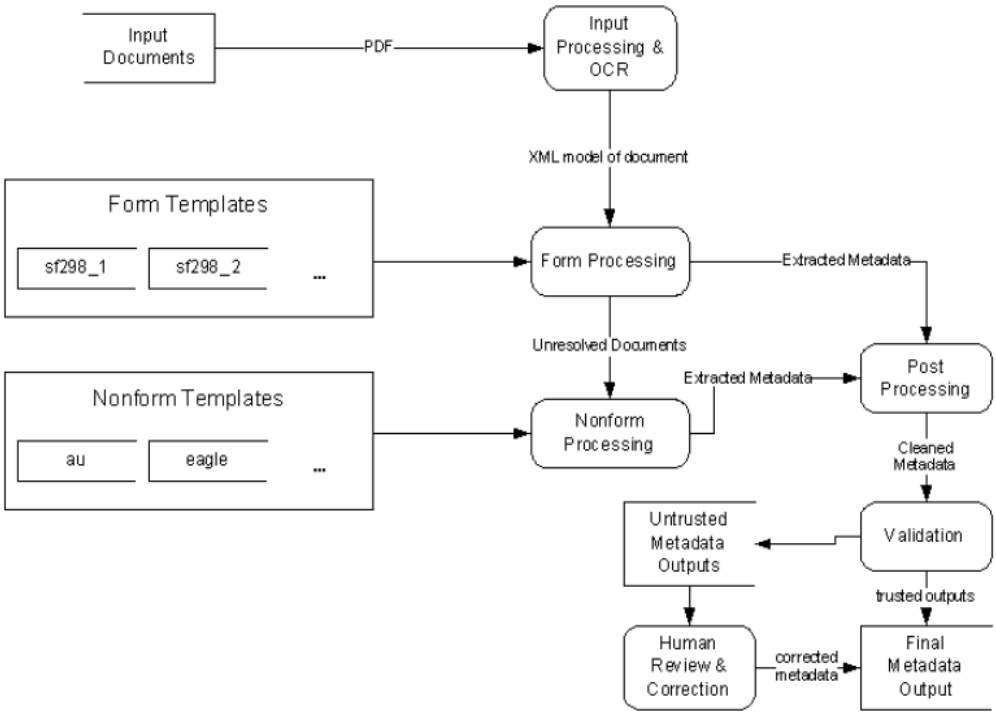

Fig. 3.8. Diagrama del Flujo de la Extracción de Metadatos Basado en Plantillas. Tomado de Figura 1 (Flynn, Zhou, Maly, Zeil, & Zubair, 2007)

El enfoque para la extracción automática de metadatos revisado en esta propuesta está basado en los *sistemas basados en reglas* (Klink, Dengel, & Kieninger, 2000) que usan

instrucciones programadas para especificar cómo extraer la información de documentos específicos. Con reglas de lenguaje suficientemente desarrolladas, tales técnicas son capaces de extraer metadatos de calidad. Sin embargo, la heterogeneidad de los OA's puede acarrear reglas complejas cuya creación y prueba llegan a consumir demasiado tiempo.

Por tal motivo, para hacer frente a los cambios por la evolución de una colección heterogénea y mantener un conjunto de reglas sin conflictos, se llevó a cabo una técnica propia de extracción de metadatos como una variante del enfoque basado en reglas, de modo que el diseño de las plantillas es simple e independiente entre sí. El listado de metadatos extraídos se muestra en la Tabla 3.9; a pesar de que la propuesta no sigue ningún estándar de metadatos en particular, la información obtenida puede ser fácilmente mapeada a DCMI y/o IEEE LOM:

Tabla 3.9. Metadatos extraídos por la Arquitectura de Extracción Basada en Plantillas.

Campo	Descripción	DCMI	LOM
Título	El nombre dado a un recurso, normalmente proporcionado por el autor.	DC.Title	1.2:General.Title
Autor	La persona u organización responsable de la creación del contenido intelectual del recurso.	DC.Creator	2.3.2:Life Cycle.Contribute.Entity cuando 2.3.1:Life Cycle.Contribute.Role tiene como valor "autor".
Palabras Clave	Expresa las claves o frases que describen el título o el contenido del recurso.	DC.Subject	1.5:General.Keyword o 9:Classification con 9.1:Classification.Purpose igual a "disciplina" o "idea".
Fecha	Una fecha en la que el recurso se puso a disposición del usuario en su forma actual.	DC.Date	2.3.3:Life Cycle.Contribute.Date cuando 2.3.1:Life Cycle.Contribute.Role tiene como valor "editor".
Resumen	Una descripción textual del recurso, como puede ser un resumen o una descripción de su contenido.	DC.Description	1.4:General.Description
Otros Colaboradores	La persona u organización que haya tenido una contribución intelectual significativa en la creación del recurso, pero cuyas contribuciones son secundarias en comparación a las de las personas u organizaciones especificadas en el elemento *Creator.*	DC.OtherContributor	2.3.2:Life Cycle.Contribute.Entity con el tipo de contribución especificada en 2.3.1:Life Cycle.Contribute.Role
Editor	La entidad responsable de hacer que el recurso se encuentre disponible en la red en su formato actual.	DC.Publisher	2.3.2:Life Cycle.Contribute.Entity cuando 2.3.1:Life Cycle.Contribute.Role tiene como valor "editor".
Tipo de Recurso	La categoría del recurso.	DC.Type	5.2:Educational.Learning Resource Type

Esta propuesta ha sido probada en la colección DTIC (Defense Technical Information Center, 2007), que contiene más de un millón de documentos PDF y a la que se suman decenas de miles de nuevos documentos cada año. Los documentos son diversos, incluyendo artículos científicos, presentaciones, tesis doctorales, actas de conferencias, folletos promocionales, las leyes públicas y leyes del Congreso. Las pruebas dieron como resultado una precisión global del 83% para los documentos con las plantillas definidas, teniendo en cuenta que en dicha colección más del 50% de los documentos tienen RDP, que contienen más de 20 campos de metadatos. Si bien, la arquitectura en si no está diseñada para ser utilizada en un RI o algún contexto educativo, esta propuesta puede ser considerada de gran utilidad dado que tiene en cuenta la heterogeneidad de los OA's y por esa misma diversidad puede llegar a ser aplicable en diversos ámbitos.

3.8 Análisis Comparativo

En la Tabla 3.10 se presenta un cuadro comparativo de las propuestas mencionadas anteriormente, a la luz de los tres aspectos relevantes en el diseño de un sistema de extracción automática de metadatos mencionados al principio de este capítulo: tipos de archivos procesados, metadatos extraídos, técnicas y recursos utilizados para el proceso de extracción.

Tabla 3.10. Cuadro comparativo propuestas de extracción automática de metadatos.

Sistema	Tipo de Archivos	Poder de Extracción de Metadatos	Técnicas y Recursos Utilizados para la Extracción Automática
SAXEF	HTML XHTML ASP PHP	No sigue un estándar de metadatos en particular, ya que produce una tarjeta propia de identificación E-learning (EIC, por sus siglas en inglés) con información definida por los desarrolladores del sistema, que permite a los profesores evaluar fácilmente cuando una página es de su interés. Extrae algunos metadatos educacionales, tales como: temas secundarios, contenido teórico o práctico, sintético o analítico, tipos y nivel multimedia, tipo de interactividad, nivel de complejidad. También extrae enlaces a otros EIC con los mismos temas o con temas relacionados.	Stop Words Reglas de mapeo directo Reglas heurísticas de mapeo Medidas estadísticas
TWYS	HTML	Extrae muchos de los metadatos de diferentes versiones del estándar LOM, entre los que se encuentran: entry, location, title, language, entity, date, format, size, description, keyword, purpose, ID. Extrae algunos metadatos educacionales tales como: interactivitytype, interactivitylevel, semanticdensity, difficulty.	Ontologías Stop Words TF/IDF (term frecuency weighting) HTML Parser Reglas de mapeo directo Reglas heurísticas de mapeo

Sistema	Tipo de Archivos	Poder de Extracción de Metadatos	Técnicas y Recursos Utilizados para la Extracción Automática
LooKIng4LO	HTML Archivos SCORM Tipos de archivos no estructurados, tales como TXT, PDF y PPT	Extrae un subconjunto menor de metadatos LOM, como author, interactivitylevel. Extrae algunos metadatos educacionales como tiempo de lectura y tiene imagen.	Ontologías Tokenizer Sentences Splitter POS Tagger Gazetteer Transducer Herramientas de procesamiento de lenguaje natural (GATE, General Architecture for Text Engineering)
MAGIC	HTML Tipos de archivos no estructurados, tales como TXT, PDF y PPT Archivos de tipo video y audio (AVI, MPG, MP3, MP4, WMA)	Extrae un subconjunto menor de metadatos LOM, entre los que se encuentran: title, keyword, entity, description. Extrae metadatos críticos conforme al estándar SCORM.	Tokenizer POS Tagger Herramientas de procesamiento de lenguaje natural (TEXTRACT)
Asistente para el Depósito de Objetos en Repositorios con Extracción Automática de Metadatos	Tipos de archivos no estructurados, tales como PDF y Word	Extrae metadatos principales conforme al estándar DCMI: title, creator, subject, description, language.	Análisis Semántico Herramientas de procesamiento de lenguaje natural (Alchemy) Análisis Sintáctico Análisis de Estructura (ParsCit)
CERMINE	PDF	No sigue un estándar de metadatos en particular, ya que el resultado está basado en la norma NISO Z39.96-2012. Sin embargo, se observa que extrae un subconjunto menor de metadatos LOM, entre los que se encuentran: title, keyword, entity, description, identifier, type.	Máquinas de Soporte Vectorial (SVM) Reglas Heurísticas
Arquitectura de Extracción Automática de Metadatos basada en Plantillas	PDF	No sigue un estándar de metadatos en particular, sin embargo, se observa que extrae un subconjunto menor de metadatos LOM, entre los que se encuentran: title, keyword, entity, description, publisher, type.	Matriz determinística y pruebas estadísticas Sistemas basados en reglas

Con base en la comparación que se presenta en la Tabla 3.10, se pueden obtener los siguientes comentarios y conclusiones:

1. En cuanto a *tipos de archivo*, las propuestas analizadas pueden trabajar, en general, con páginas web HTMLy archivos PDF, debido, en primer lugar, a la gran cantidad de estos recursos de aprendizaje que se encuentra disponibles en Internet y, en segundo lugar, porque las

páginas web HTML tienen etiquetas que poseen más información acerca de su contenido y los archivos PDF pueden ser convertidos en formatos XML para su procesamiento. De manera particular, los archivos SCORM, que son estructurados y tienen algunos campos de metadatos ya clasificados, sólo son tratados por Looking4LO.

Algunos de los tipos de archivos no estructurados, que no presentan etiquetas o metadatos, tales como TXT, DOC, DOCX y PPT, son tratados por Looking4LO, MAGIC y el Asistente. Finalmente, MAGIC es el único sistema capaz de procesar tipos de OA interesantes, como lo son los archivos de video y audio. En general, de las propuestas revisadas, MAGIC es la que contempla una gran cantidad de tipos de archivos comúnmente utilizados.

2. En lo que respecta a los metadatos extraídos, tres de las propuestas revisadas extraen metadatos del estándar LOM, SAXEF produce una tarjeta propia de identificación E-learning (EIC), el Asistente extrae metadatos del estándar DCMI y los demás no mencionan ningún estándar de metadatos. TWYS es el sistema que mayor cantidad de metadatos extrae (tanto generales como educativos), siguiendo el estándar LOM. CERMINE y la Arquitectura Basada en Plantillas no realizan la extracción bajo ningún estándar, sin embargo, los metadatos extraídos pueden ser fácilmente correspondidos a los estándares LOM y/o DCMI.

Se observa también que todas las propuestas extraen los siguiente metadatos generales: título, palabras clave, autor, tipo de recurso y resumen; esto constituye una buena partida para los propósitos de búsqueda de OA en RI.

3. Con respecto a la extracción de metadatos educacionales, TWYS genera cuatro metadatos educacionales, Looking4LO solamente extrae uno y MAGIC no extrae ninguno. En cuanto a SAXEF, si bien esta no utiliza el estándar LOM, en la tarjeta resultado hay metadatos que identifican si una página web es teórica o práctica y sintética o analítica. Las demás propuestas no hacen énfasis en la extracción de metadatos educacionales. Estos metadatos son de gran importancia para la identificación y recuperación de cualquier OA, ya que suministran información de interés acerca del contenido educativo del mismo y pueden apoyar los sistemas de búsqueda recomendadores, en cuanto al ajuste de preferencias de acuerdo al perfil del usuario (profesor o estudiante). Sin embargo son uno de los metadatos más difíciles de obtener en forma automática.

4. Por último, se mencionan las técnicas y recursos utilizados para la extracción automática de metadatos. En general, para el proceso de extracción automática, cada propuesta hace uso de

más de un recurso de procesamiento, entre los que se encuentran Stop Words, Ontologías, Tokenizer, reglas de mapeo directo, reglas heurísticas de mapeo, entre otros, teniendo en cuenta que cada uno de estos recursos cumple una función determinada dentro del proceso de extracción y que, parte de la información encontrada en los OA, puede ser traducida directamente a algunos metadatos. Entre tanto, otros metadatos requerirán reglas o métodos no triviales para determinar su valor.

Teniendo en cuenta lo expuesto anteriormente para las diferentes propuestas de extracción automática de metadatos, se puede deducir que:

a) No se contemplan gran cantidad de formatos de archivos; esto puede llegar a limitar la funcionalidad y utilidad del repositorio.

b) No se extraen de manera significativa metadatos educacionales, que resultan ser sumamente importantes a la hora de la recuperación de los OA mediante ejecución de búsquedas.

c) En la mayoría de los casos se sigue únicamente el estándar de metadatos LOM, lo que puede limitar la calidad de la información descriptiva asociada a los OA y su integración con repositorios institucionales dado que estos, en su mayoría, hacen uso del estándar de metadatos DCMI.

d) Si bien actualmente hay algunos estudios y sistemas para la extracción automática de metadatos, falta mucho por hacer, en parte, porque implica la aplicación de estrategias de inteligencia artificial.

e) Estos extractores no están pensados para ser usados en repositorios institucionales si no para propósitos generales.

f) La mayoría de las propuestas revisadas son el resultado de trabajos de investigación académica, por lo que están implementadas como aplicaciones web, se encuentran disponibles gratuitamente y cuentan con documentación para facilitar su uso.

Para el diseño del extractor que se propone en esta tesis se consideraron las comunidades, colecciones y metadatos definidas en (Giorgetti, Romero, & Gutierrez, 2015) donde se tuvieron en cuenta distintos estándares de metadatos los cuales aparecen identificados como obligatorios y opcionales dependiendo de cada colección.

En dicho resumen se puede observar lo siguiente:

a) Se consideraron en total 19 metadatos y 13 colecciones.
b) Los metadatos considerados están contemplados dentro de los estándares DCMI e IEEE LOM.
c) Se establecieron equivalencias entre metadatos pertenecientes a los diferentes estándares considerados. Ejemplo: contributor/lifecycle.contribute corresponde a: Colaborador/Sponsor; Fuente de Financiamiento, Director u Organizador.
d) Los siguientes metadatos son comunes para todas las colecciones: fecha de publicación, resumen, filiación, contexto, dificultad, tipo de interactividad, rango de edad, formato, idioma, relación, fuente, palabra clave, título y tipo.

Capítulo

4

AMELOIR: Algoritmo para Extracción de Metadatos

Teniendo en cuenta que en el capítulo 3 se analizaron y revisaron las ventajas y características principales de los trabajos relacionados con esta tesis, en este capítulo se presenta la propuesta del nuevo algoritmo para la extracción automática de metadatos AMELOIR (Automatic Metadata Extraction Learning Object Institutional Repository).

AMELOIR fue diseñado principalmente para ser utilizado en RIs de acceso abierto y que puede ser integrado en entornos académicos en los cuales se haga uso de DSpace, tal como sugieren San Martin y otros (San Martin, Guarnieri, & Bongiovani, 2014). La utilidad de AMELOIR en entornos no académicos está sujeta a la revisión de la lista de colecciones que se presenta durante el proceso de carga del OA, ya que esto se convierte en el punto de partida del enfoque del proceso de extracción.

4.1 Requisitos Funcionales y No Funcionales

AMELOIR cumple con los siguientes requisitos funcionales y no funcionales teniendo en cuenta que realiza una función muy específica dentro del proceso de carga de un documento al repositorio institucional, interactuando con el software DSpace, de manera transparente para el usuario.

Requisitos Funcionales

1. Debe extraer e inferir la información de los metadatos correspondientes al documento sometido al proceso de carga en DSpace.

2. Debe poder funcionar con archivos en los formatos más populares: PDF y Microsoft Word (.DOC y .DOCX).

3. No soportará la extracción de información de documentos que provengan de fuentes escaneadas (.JPG).

4. La extracción de metadatos se debe ejecutar de manera automática durante el proceso de carga de un documento al software del repositorio, sin interacción con el usuario.

Requisitos No Funcionales

1. Los metadatos extraídos deben ser compatibles con los estándares Dublin Core e IEEE LOM (presentados en la sección 2.2).

2. El tiempo que demanda la extracción de metadatos no debe ser superior a 60 segundos, contados después de ubicar el archivo en la máquina local, seleccionar el tipo de recurso al que pertenece el documento y dar click en "Siguiente", tras lo cual el algoritmo debe devolver la información extraída hasta ese momento y retornar el control al proceso de carga de documentos.

3. Tanto el código del algoritmo como las dependencias que utilice el extractor, deben corresponder a licencias de código abierto y/o gratuito.

4.2 Funcionamiento del Extractor

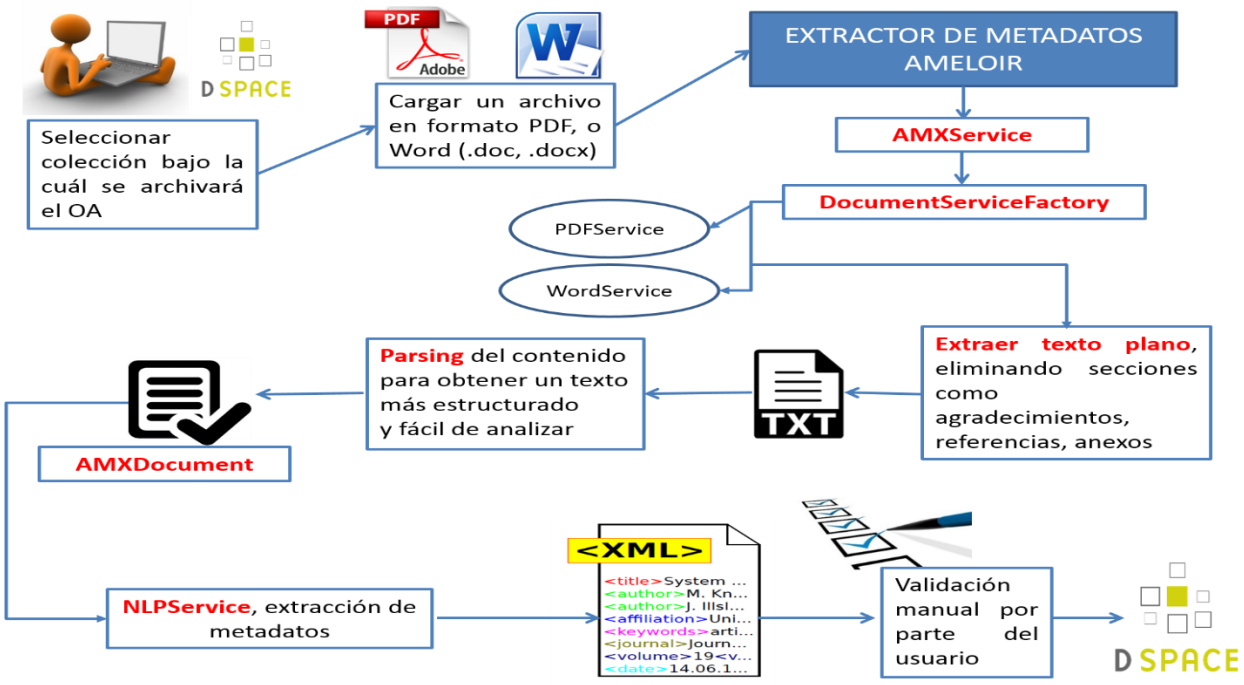

Fig. 4.1. AMELOIR algoritmo propuesto para la extracción de metadatos.

Tal como muestra la Figura 4.1, el proceso de extracción inicia cuando el usuario selecciona la colección bajo la cual archivará el OA en el Repositorio Institucional de Acceso Abierto. El siguiente paso, consiste en que el usuario carga el archivo que desea almacenar en el repositorio. Inicialmente, el algoritmo será implementado para trabajar con archivos en

formato PDF y Word (.doc, .docx), pero se contempla como trabajo futuro ampliar la propuesta para trabajar con más tipos de archivos. Estos archivos serán enviados al extractor, para dar comienzo al proceso de extracción automática de metadatos.

El extractor se comunica con el RI a través de un único método definido en la clase principal del sistema, AMXService. Este método devuelve como resultado una clase con los metadatos extraídos, la cual luego es mapeada a un archivo XML. Internamente, AMXService se comunica con 3 servicios principales, representados por las siguientes interfaces:

a) *IDocumentService*: encargado de extraer el texto del documento, remover secciones que no son consideradas necesarias para la extracción de los metadatos (por ejemplo: agradecimientos, código/pseudocódigo, anexos, referencias) y parsear el contenido para obtener un texto más estructurado y fácil de analizar. Una parte importante de este servicio es la clase *DocumentServiceFactory*, la cual define qué implementación se utilizará de acuerdo al tipo de archivo que está siendo analizado.

   Para cada implementación se utilizan diferentes librerías, Apache POI y PDFxStream, las cuales ofrecen métodos especializados, no sólo para extraer el texto del documento, sino también para analizar sus propiedades y formato.

b) *IBooksService*: es un servicio que intenta obtener los metadatos desde las bases de datos de Google Books y Worldcat, antes de inferirlos del texto. Busca previamente en el texto un identificador (como ISBN, ISSN o en su defecto, el título del documento) y realiza una consulta para obtener los metadatos.

c) *INLPService*: ofrece una serie de métodos para obtener información del texto utilizando una combinación de técnicas de NLP y llamadas a librerías NLP. Hace uso de los servicios *AlchemyService*, *CoreNLPService*, *KeywordsService*, los cuales fueron definidos para tener una mayor abstracción de las librerías que utilizan AlchemyAPI, Apache CoreNLP y Lucene/Freeling, respectivamente.

Cuando se hace una llamada al método de extracción, *AMXService* pide a *DocumentServiceFactory* el servicio necesario para manejar el documento. De esta forma, se obtiene la instancia de la implementación correspondiente de *IDocumentService*, la cual dependerá del formato del archivo (PDFService y WordService para documentos en formato PDF y .doc/.docx respectivamente). Luego, se utiliza este servicio de extracción de texto para

generar la entidad *AMXDocument*, la cual es una representación estructurada del documento, con toda la información extraída del mismo. Este resultado se utilizará como entrada para las llamadas a los demás servicios.

El siguiente paso es utilizar el servicio *NLPService* para extraer el metadato *identifier* que se utilizará para hacer una consulta a *BooksService* e intentar obtener los metadatos desde Google Books (si el tipo de recurso es un Book) o Worldcat (si el tipo de recurso es un Article). Si el paso anterior no encuentra información sobre el recurso, se extraen los metadatos *author* y *title* para intentar hacer una nueva consulta con estos valores. La extracción de los metadatos anteriores se realiza llamando a métodos de los servicios *NLPService* y *DocumentService*, respectivamente.

El metadato *title* es el único que se obtiene desde el documento original y no desde la abstracción del mismo, debido a que para su extracción se requiere información sobre el formato del texto, el cual no es guardado en la clase.

Finalmente, se consulta a *NLPService* para extraer todos los metadatos que aún no tengan asignado un valor, ya sea porque *BooksService* no encontró datos sobre el documento, o porque sólo devolvió una parte de ellos. En la Figura 4.2 se presenta un diagrama de componentes que muestra las distintas librerías y componentes del extractor de metadatos AMELOIR.

Además de estas interfaces, se cuenta con una serie de módulos *Helpers* definidos para facilitar la realización de ciertas funcionalidades comunes. Estos módulos buscan encapsular tareas similares que se utilizan con frecuencia dentro de los servicios:

- *RegexHelper*: el más complejo de los módulos debido a la gran necesidad de análisis de expresiones regulares en el texto. Contiene una gran cantidad de métodos que permiten hacer suposiciones acerca de la sección en la que se encuentra la expresión dentro del documento, y a partir de esta información, obtener más detalles aplicando expresiones más complejas. También es utilizada durante el parsing del documento, para excluir las secciones que no son de utilidad para la extracción de metadatos (previamente mencionadas).

- *DBHelper*: permite hacer consultas a la base de datos.

- *TextHelper*: contiene funcionalidades básicas de análisis de texto y parsing, que permiten estructurar el texto.

- *ComparatorHelper*: contiene funcionalidades para realizar comparaciones entre textos, como son: obtener la distancia de Levenshtein entre dos palabras, definir si dos palabras son similares, definir si determinado texto o una versión similar al mismo está incluido en una lista, entre otras.

En la Figura 4.3 se muestra el flujo de carga por defecto en DSpace. Los pasos en color verde corresponden a la carga del archivo en DSpace. El paso de color naranja corresponde a la definición de metadatos y los pasos en color rojo hacen referencia a la verificación de información y finalización del proceso de carga.

- *Selección de la colección*: el usuario debe especificar la colección en la que desea almacenar su OA. Para ello, se presenta una lista desplegable (etiquetada como Resource Type) en el que se mostrarán todas las opciones de las colecciones posibles.

- *Descripción del elemento*: en este paso se presentan los campos necesarios para que se describa en detalle el OA. En ellos se introduce la información más relevante del ítem, información a través de la que más tarde se podrán hacer las búsquedas para recuperar el objeto.

- *Subida del archivo*: después de describir el elemento con sus respectivos metadatos, el sistema solicitará que se cargue el o los archivos deseados, con la posibilidad de ingresar una descripción de los mismos.

- *Verificación del envío*: se muestra una visión general del ítem con las secciones de los metadatos, tal como se han ingresado, en cada una de las cuales se puede realizar cambios. Se detallan también la colección en la que será depositado, los nombres de los ficheros, etc.

- *Autorización y licencias*: en este paso el usuario debe leer los términos y condiciones del repositorio, y aceptarlos o rechazarlos. La licencia que se muestra por defecto indica que el autor da la libertad a DSpace de hacer respaldos del ítem con fines de conservación, y que los derechos de autor serán respetados.

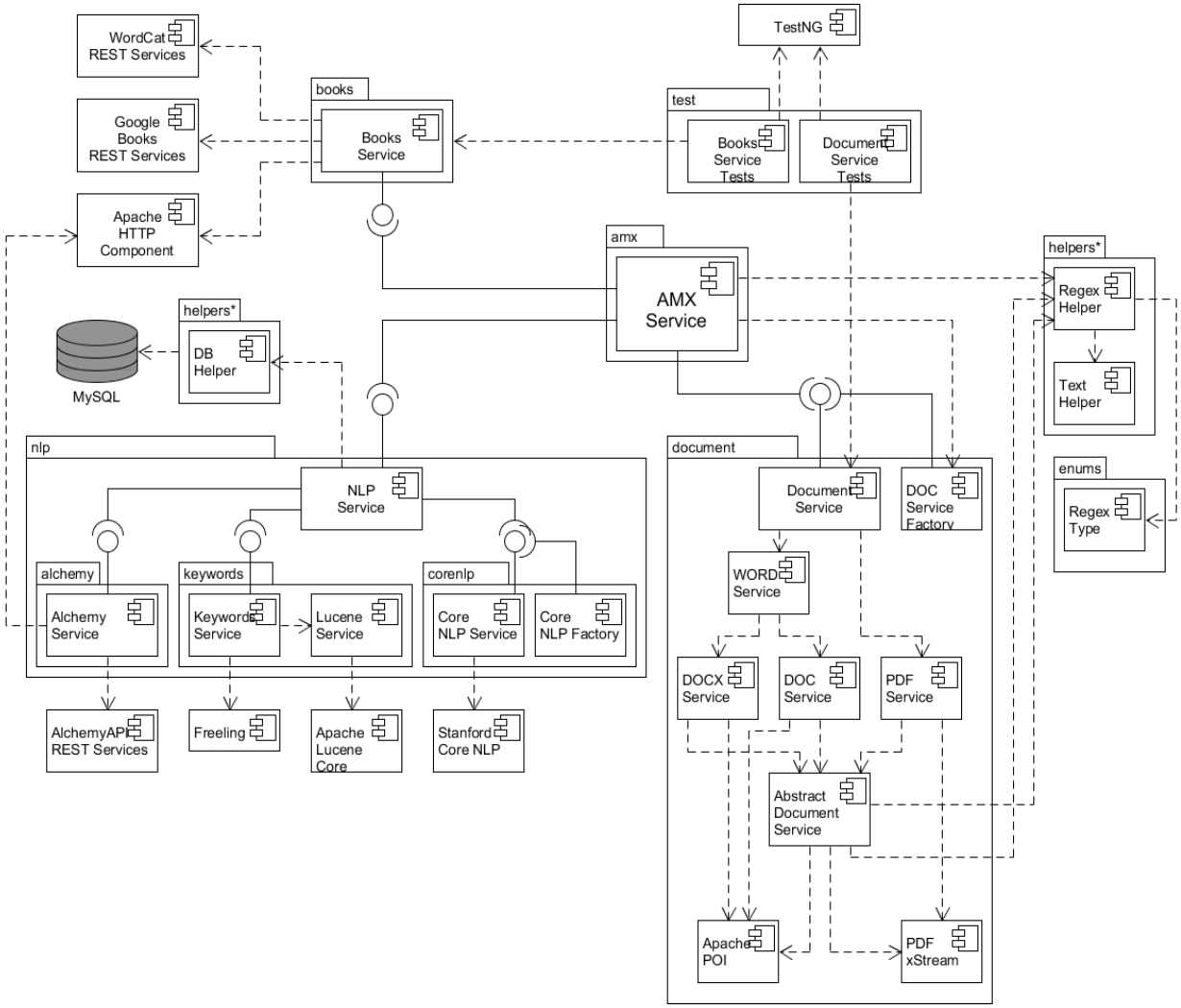

Fig. 4.2. Diagrama de componentes AMELOIR.

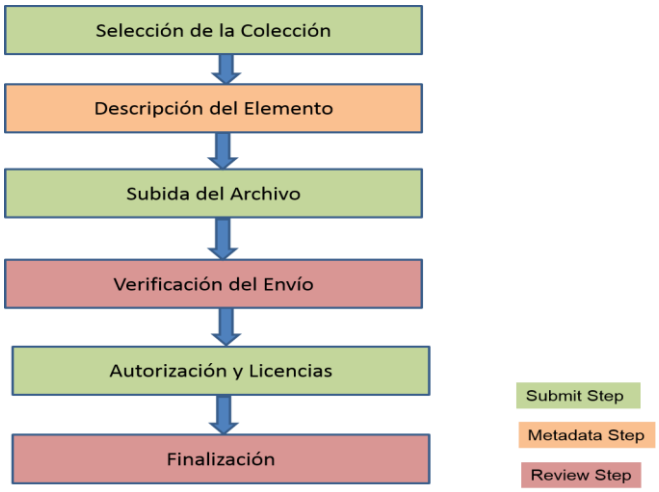

Fig. 4.3. Flujo de carga DSpace. Adaptado de Figura 1 (Casali, Deco, Bender, Fontanarrosa, & Sabater, 2013)

Con el fin de focalizar los esfuerzos de investigación y desarrollo del algoritmo de extracción AMELOIR, se decidió modificar el proceso de carga por defecto descrito anteriormente con el objetivo de acotarlo, reorganizando los pasos e introduciendo como valor agregado el proceso de extracción automática de metadatos.

El proceso de carga diseñado para la propuesta consiste en cinco pasos, tal como se muestra en la Figura 4.4, uno de los cuales es opcional (marcado con *), para dar la posibilidad al usuario de cumplir el objetivo de almacenar el OA junto con sus metadatos, en el caso en que el extractor de metadatos halle menos o ninguno de los metadatos requeridos para el almacenamiento del objeto.

- *Subida del archivo*: en primer lugar se selecciona el archivo que se desea subir a DSpace y que representa el OA.
- *Selección de colección*: para continuar con el proceso, debe seleccionarse el tipo de recurso al que pertenece el archivo; este paso permanece inalterado con respecto al flujo de carga por defecto de DSpace.
- *Ejecución del algoritmo de extracción AMELOIR*: el siguiente paso se ejecuta en segundo plano, es decir, sin mostrar una interfaz al usuario. Es en este paso donde se invoca al extractor de metadatos AMELOIR para que procese el archivo subido.

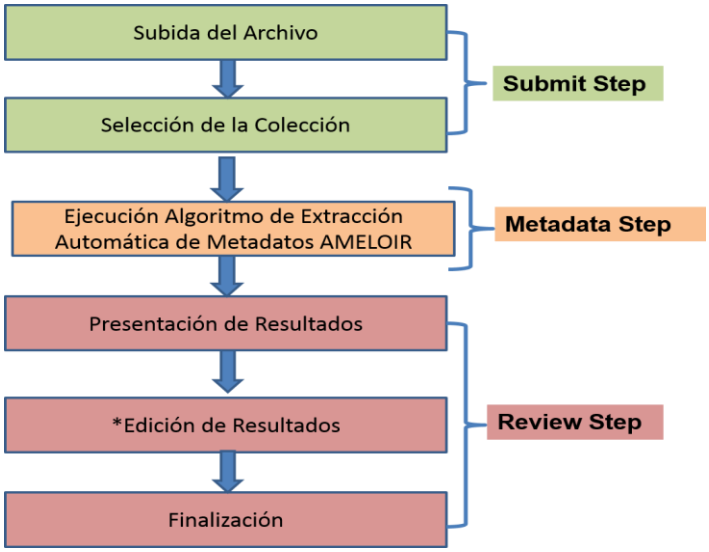

Fig. 4.4. Incorporación del algoritmo propuesto AMELOIR al flujo de carga DSpace.

- *Presentación de resultados del extractor de metadatos*: aquí se muestran los resultados obtenidos del procesamiento realizado en el paso anterior, dependiendo de cuáles fueron los metadatos hallados por el algoritmo de extracción.

- **Edición de resultados del extractor de metadatos:* este es un paso opcional que se lleva a cabo por decisión del usuario que está archivando el OA en el RI. Por ejemplo, en el caso en que el extractor de metadatos halle menos o ninguno de los metadatos requeridos para el almacenamiento del objeto, o que complete en forma errónea algún metadato, el usuario pueda igual cumplir con el objetivo de almacenar el OA con sus metadatos asociados. Para ello, se provee un enlace directo a la sección de edición del ítem (Go to the "Edit item" section).

- *Finalización del proceso:* una vez que se completa el envío, los datos son almacenados junto al ítem subido.

4.3 Colecciones y Metadatos

En el RI se van a cargar documentos de diversas fuentes, por lo que los mismos no van a contar con un diseño homogéneo. De acuerdo a las propuestas de diseño que presentan San Martin y otros (San Martin, Guarnieri, & Bongiovani, 2014), se establecieron las siguientes 6 colecciones predeterminadas a modo de categorías, que responden a las distintas producciones de OA de la comunidad académica:

- Artículo de revista.
- Capítulo de libro.
- Libro.
- Tesis (grado y posgrado).
- Paper.
- Reporte técnico.

La diversidad de OA afectó en gran medida la forma en que fueron programados los algoritmos de extracción para cada metadato requerido. Es imposible tener en cuenta la infinidad de formas de redactar artículos de revista, tesis, papers, etc., por lo que se hicieron generalizaciones y estimaciones arbitrarias para cada metadato, tratando de obtener los mejores resultados posibles para la mayor cantidad de documentos tomados como conjunto

de prueba. En general, los algoritmos se basaron en el conocimiento previo del formateo más común para los bloques de cada documento, incorporando además, el uso de información textual (patrones de texto), información de la tipografía (tamaño), entre otros atributos (tamaño y cantidad de párrafos, por ejemplo) (Klink, Dengel, & Kieninger, 2000).

Teniendo en cuenta el trabajo interdisciplinario realizado en el proceso de definición de las diversas colecciones y sus respectivos metadatos obligatorios y optativos, en las Tablas 4.1 y 4.2 se presentan dichos metadatos, bajo los estándares de metadatos Dublin Core e IEEE LOM. La columna “multivaluado” hace referencia a si el metadato puede tener más de un valor, como por ejemplo autor, ya que un documento puede tener más de un autor.

4.3.1 Metadatos Obligatorios

A continuación se detallan los algoritmos utilizados para extraer los metadatos obligatorios referenciados en la Tabla 4.1., presentando para cada uno de ellos una descripción y su correspondiente diagrama de flujo:

Tabla 4.1. Metadatos Descriptivos Obligatorios.

Metadato DCMI	Metadado IEE LOM	Multivaluado	Etiqueta
dc.title	1.2:General.Title	SI	Título
dc.creator	2.3.2:Life Cycle.Contribute.Entity cuando 2.3.1:Life Cycle.Contribute.Role tiene como valor "autor".	SI	Autor
dc.description	1.4:General.Description	SI	Descripción
dc.subject	1.5:General.Keyword o 9:Classification con 9.1:Classification.Purpose igual a "disciplina" o "idea".	SI	Palabras clave
dc.language.iso	1.3:General.Language	NO	Idioma
dc.type	5.2:Educational.Learning Resource Type	NO	Tipo de Recurso
dc.publisher	2.3.2:Life Cycle.Contribute.Entity cuando 2.3.1:Life Cycle.Contribute.Role tiene como valor "editor".	NO	Editorial
dc.date.issued	2.3.3:Life Cycle.Contribute.Date cuando 2.3.1:Life Cycle.Contribute.Role tiene como valor "editor".	NO	Fecha de publicación
dc.format	4.1:Technical.Format	NO	Formato
dc.identifier	1.1.2:General.Identifier.Entry	SI	Identificador

Tabla 4.2. Metadatos Descriptivos Opcionales.

Metadato DCMI	Metadato IEELOM	Multivaluado	Etiqueta
dc.contributor.editor dc.contributor.illustrator dc.contributor.other	2.3.2:Life Cycle.Contribute.Entity con el tipo de contribución especificada en 2.3.1:Life Cycle.Contribute.Role	SI	Colaborador(es) en Edición, Ilustración u Otros
	5.6 Educational.Context	SI	Contexto
dc.description.fil		SI	Filiación
dc.rights	6.3:Rights.Description	NO	Derechos
	5.7 Educational.Typical Age Range	SI	Rango de edad
	5.8 Educational.Difficulty	SI	Audiencia & Dificultad

Título y Subtítulo

Para el análisis de este metadato se consideraron dos posibilidades: el estilo más popular para destacar un título es asignarle el mayor tamaño de fuente, normalmente en las primeras 3 páginas, o bien que se encuentre encerrado entre comillas. Fueron implementados dos métodos para buscar candidatos a títulos siguiendo estas directivas. Para considerar un resultado como candidato, se verifica que posea una cantidad de palabras mínima y máxima (mínimo 2 y máximo 60 palabras), que consista mayoritariamente de caracteres alfabéticos, que no posea determinadas palabras referidas a instituciones o títulos de carrera y, por último, que no pertenezca al encabezado de página.

En base a lo encontrado por cada método, se define el de mayor longitud como título y, si lo hay, el de menor longitud como subtítulo. Rara vez los documentos poseen subtítulos, sin embargo, de esta manera aumentan las posibilidades de que el título correcto se muestre como metadato extraído. En la Figura 4.5 se muestra el diagrama de flujo que representa el algoritmo de extracción del metadato título:

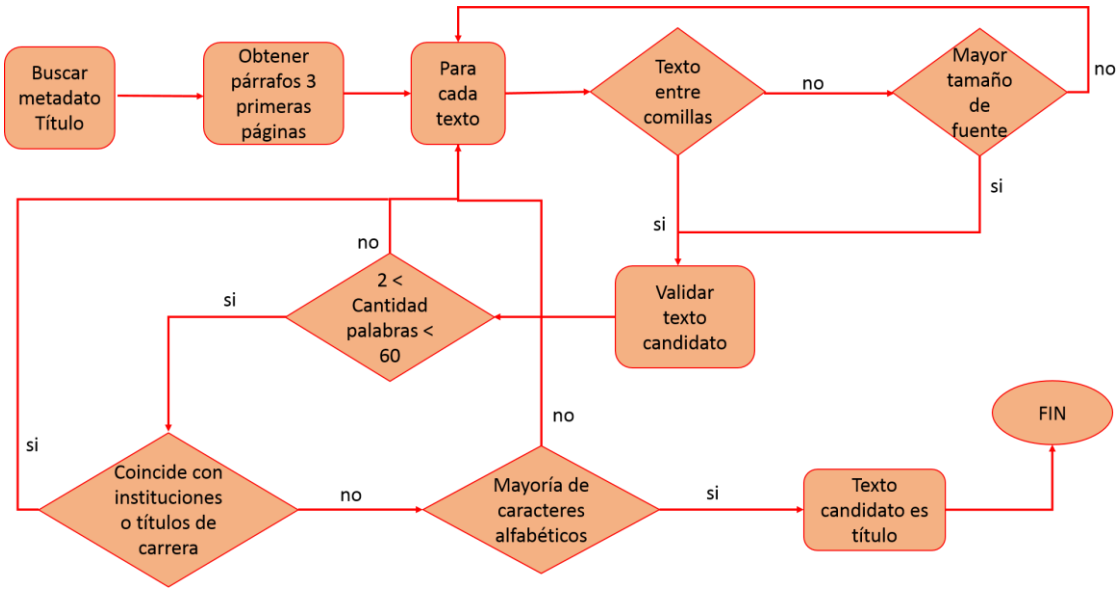

Fig. 4.5. Diagrama de flujo algoritmo de extracción metadato Título

Autor

Inicialmente, se define la parte del documento a ser analizado, debido a que los nombres de los autores suelen encontrarse en las primeras 3 hojas del mismo. Se toma como texto a analizar todo lo que se encuentre en dichas páginas, delimitando la misma a partir del inicio del documento hasta el punto en que se identifiquen alguna sección que indique que inicia el contenido (por ejemplo, introducción, índice, resumen, entre otros). En caso de no lograr definir el inicio del contenido, se toma un número limitado de párrafos como referencia, que para el caso de AMELOIR corresponde a un total de 30 párrafos. A dichos párrafos se les denominó "*párrafos de interés*".

Tomando como base estos *párrafos de interés*, se realiza una identificación de todos los nombres propios incluidos en el mismo. Los candidatos se definen a partir de técnicas de NER (Named Entity Recognition): inicialmente se toman las palabras capitalizadas consecutivas (aquellas cuya primera letra se encuentra en mayúscula) y las palabras en mayúsculas consecutivas; hay que tener en cuenta varios casos en los que no se da estrictamente esta regla, como en nombres compuestos (ej. "María de los Ángeles"), o simplemente nombres escritos completamente en mayúsculas. Para ello, los nombres identificados se normalizan, primero reemplazando los caracteres '-' por espacios en blanco, luego reemplazando los dígitos del 0 al 9 que puedan existir por la cadena vacía, y por último, reemplazando los conectores de nombres compuestos (de|del|van|von|la|los|den) por espacios en blanco.

Como parte de la implementación de AMELOIR se creó una base de datos en MySQL que contiene una serie de tablas de apoyo para el proceso de extracción de los metadatos; cabe aclarar que el mantenimiento de esta base de datos será responsabilidad del administrador del repositorio institucional en el que se implemente el extractor. Para la extracción del metadato *Author* se tienen en cuenta dos tablas: una que contiene nombres propios y otra que contiene apellidos; ambas tablas cuentan con un identificador numérico como llave primaria, el dato (name o surname) y el lenguaje del dato (es: Español, en: Inglés). La tabla de apellidos tiene alrededor de 113.000 registros, mientras que la tabla de nombres propios cuenta con alrededor de 14.600 registros.

Para aumentar la confiabilidad del resultado, se verifican los candidatos contra la base de datos de nombres propios y apellidos y se les da un puntaje al conjunto de palabras para saber si califica como nombre completo. En esta puntuación influyen otros factores para evitar que se identifiquen "falsos nombres", como por ejemplo la verificación de palabras comunes, es decir, aquellas palabras que pueden ser tanto nombres de personas como de ciudades (ej. "Rosario"), nombres que coinciden con palabras comunes del idioma (ej. En inglés, "Will" o en español, "Norma"), entre otras; si el candidato contiene palabras comunes se le asigna un puntaje de 0.5, de lo contrario, se le asigna un puntaje de 1.

El puntaje total del candidato estará determinado por la división del puntaje total obtenido entre la cantidad de palabras del candidato; si el resultado de esta división es superior a 0.5, el candidato se añade al listado de metadatos de *Author*.

Finalmente, se eliminan duplicados de la lista encontrada. Por lo general, los nombres de autores y contribuyentes se repiten en el texto, y no siempre en el mismo formato, por ejemplo se debe determinar que los nombres "María Ana Pérez", "Ma. Ana Pérez" y "Pérez María Ana" se refieren a la misma persona.

En la Figura 4.6 se muestra el diagrama de flujo que representa el algoritmo de extracción del metadato autor:

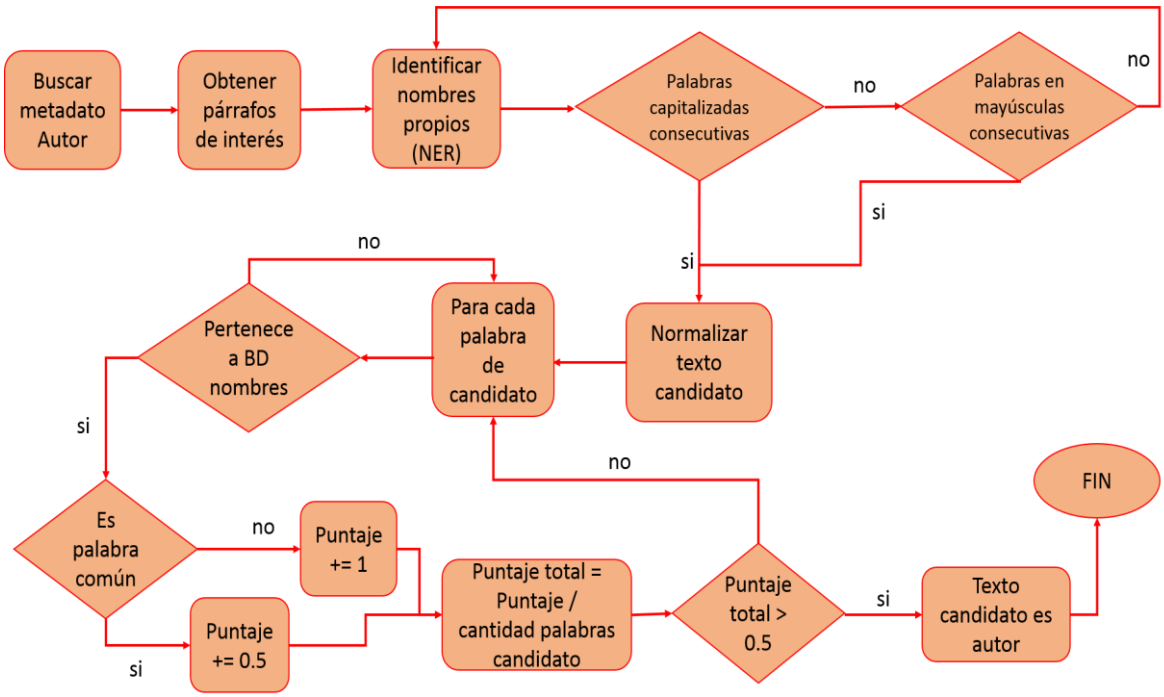

Fig. 4.6. Diagrama de flujo algoritmo de extracción metadato Autor

Descripción

Normalmente, los documentos incluyen un breve párrafo que resume lo tratado en el resto del contenido. En primer lugar se busca textualmente el título de la sección *Resumen*, y se pasa el texto del párrafo como el metadato correspondiente a *Description*.

En el caso que no se encuentre presente la sección *Resumen*, se toma el texto correspondiente a alguna de las secciones informativas del documento (introducción, conclusiones, contenido, objetivos, prefacio, prólogo). En la Figura 4.7 se muestra el diagrama de flujo que representa el algoritmo de extracción del metadato descripción:

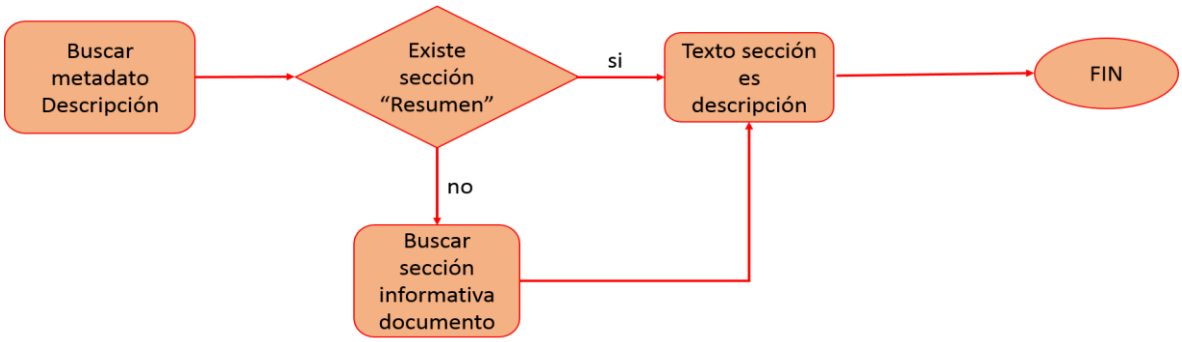

Fig. 4.7. Diagrama de flujo algoritmo de extracción metadato Descripción

Palabras Clave

Por definición, las palabras claves describen los principales tópicos expresados en un documento. Por lo tanto, es importante que el repositorio cuente con este metadato para poder vincular los documentos relacionados. Para algunos tipos de documentos, como por ejemplo papers y artículos de revista, es común que este dato esté presente explícitamente. Por consiguiente, lo primero que se intenta efectuar con respecto a este metadato es la extracción directa, buscando una mención al mismo en el texto de las primeras 3 páginas del documento. De no ser exitosa esta búsqueda, se emplea un algoritmo de análisis de palabras clave para obtenerlas.

Tal como muestra la Figura 4.8, un algoritmo típico de extracción de palabras clave tiene tres componentes principales (Berry & Kog, 2010):

- *Selección de candidatos*: se extraen todas las posibles palabras, frases, términos o conceptos (dependiendo de la implementación) que pueden ser potencialmente palabras clave.

- *Cálculo de propiedades*: para cada candidato, se calculan las propiedades que indican que puede ser una palabra clave. Por ejemplo, un candidato que aparece en el título es probable que sea una palabra clave.

- *Puntuación y selección de palabras clave*: todos los candidatos pueden ser puntuados, ya sea combinando sus propiedades con una fórmula, o usando una técnica de aprendizaje automático para determinar la probabilidad de un candidato a ser palabra clave. Un determinado umbral de puntaje o probabilidad, o un límite en la cantidad de palabras clave es usado para seleccionar el conjunto final de palabras clave.

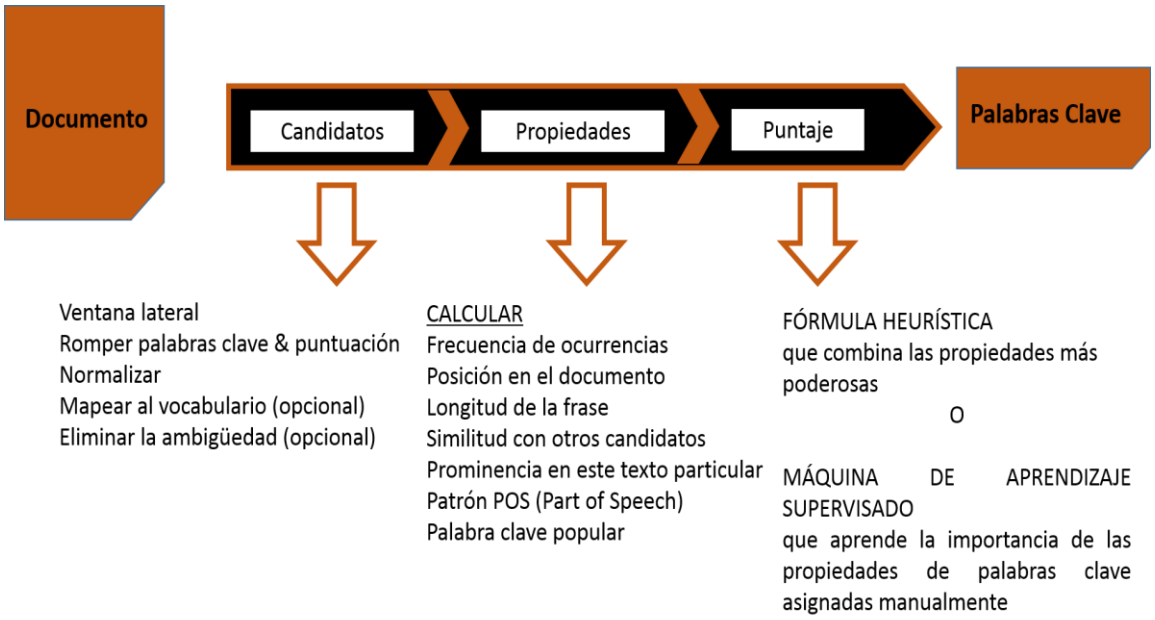

Fig. 4.8. Algoritmo típico de extracción de palabras claves. Traducción figura original (Berry & Kog, 2010)

Tomando de base este algoritmo general, fue implementado uno propio a partir de las herramientas brindadas por las librerías utilizadas.

Primero, se analizan las palabras del documento para categorizarlas con POS Tagging, una de las técnicas de procesamiento de lenguaje natural introducida en la sección 2.3.4, de modo de obtener el lemma (la forma canónica o cita de una palabra) de cada una y su categoría morfológica. Luego se limpia el listado de las palabras analizadas, descartando aquellas que sean Stop Words y que no correspondan a un sustantivo.

Posteriormente, se calcula la frecuencia de cada palabra en base a su lemma, de las cuales se tienen en cuenta las N palabras de mayor frecuencia para continuar el proceso de selección, siendo N un número arbitrariamente elegido para descartar las palabras menos frecuentes de los posteriores cálculos; para el caso de AMELOIR N = 40.

Por último, para cada candidato *i* a palabra clave de este conjunto, se efectúa un cálculo de puntuación. Para asignar el puntaje, se diseñó la fórmula (1):

$$Puntaje_i = F_i + T_i + P_i \quad (1)$$

Donde, **F** es la frecuencia de aparición del candidato en el documento, **T** es un valor que indica si el candidato es parte o no del título, y **P** es la posición de la primera mención del candidato

en el documento. A continuación se detallará el cálculo de cada uno de los componentes de la ecuación.

Para calcular **F** se asigna un puntaje de referencia (*frecuencia relativa*) entre la frecuencia de cada candidato i-esimo con respecto a la cantidad de palabras totales del texto según se indica en la fórmula (2):

$$Frec.Relativa_i = \frac{Frecuencia_i^2}{Total\ Palabras} \tag{2}$$

Luego, se toma el valor más alto de frecuencia relativa (el denominador corresponde a la mayor frecuencia relativa de todos los candidatos) y se normaliza el puntaje de cada uno de los candidatos en relación a este valor (fórmula 3), formando el primer componente de la fórmula (1):

$$F_i = \frac{Frec.Relativa_i}{Max(Frec.Relativa_x)} * 30 \tag{3}$$

El componente $\mathbf{T_i}$ de la fórmula (1) hace referencia al hecho de que el candidato en cuestión esté presente en el título. Es el término de mayor peso, y resulta:

$T_i = 50$, si el candidato *i* está en el título.

$T_i = 0$, en caso contrario.

El último componente de la fórmula (1) es el puntaje **P**, el cual refleja la posición de la primera mención del candidato en el documento. El valor de **P** se asigna según las siguientes consideraciones:

$P_i = 20$, si el candidato *i* se menciona por primera vez dentro del primer cuarto del documento.

$P_i = 10$, si la primera mención del candidato *i* es después del primer cuarto del documento y antes de la mitad.

$P_i = 5$, si el candidato *i* es mencionado en el último cuarto.

Esta fórmula resulta en un valor máximo de 100, siendo éste el caso de la palabra más frecuente del documento que forme parte del título, y que sea mencionada en las primeras

secciones del documento. El diagrama de flujo de la Figura 4.9 muestra los pasos a seguir para determinar las palabras claves.

Una vez hechos los cálculos para todos los candidatos, se seleccionan como palabras claves del documento las correspondientes a los 8 mayores puntajes.

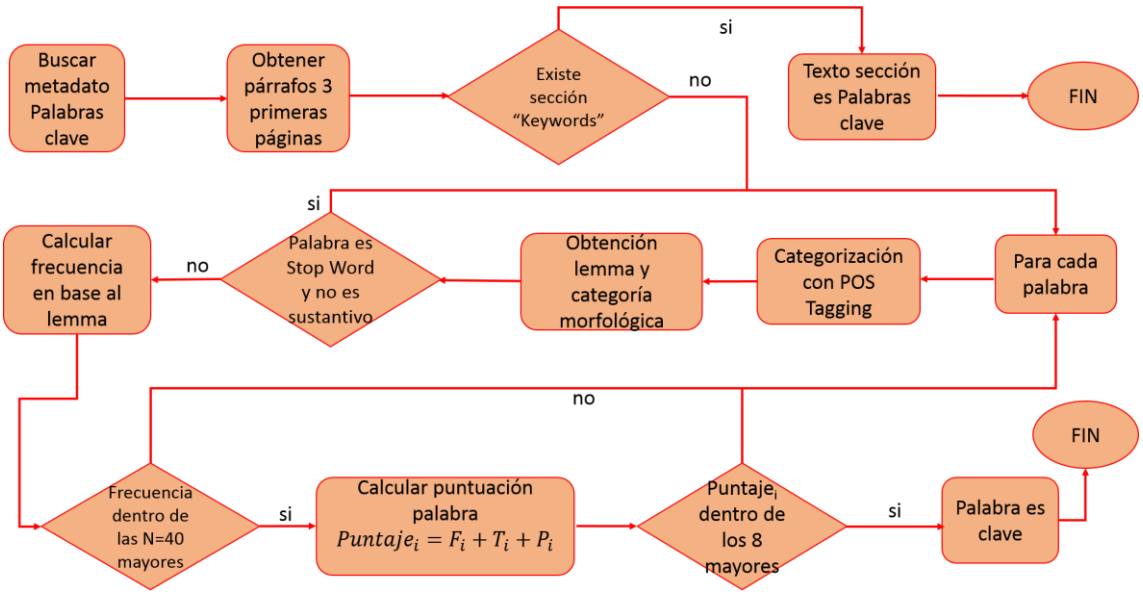

Fig. 4.9. Diagrama de flujo algoritmo de extracción metadato Palabras Clave

Idioma

Para el caso de AMELOIR se trabaja con OAs en idioma inglés y español; se evaluaron varias librerías para la extracción de este metadato y finalmente se decidió utilizar *AlchemyAPI*, debido a que era la que obtenía resultados más exactos. La principal debilidad de las demás librerías era confundir el idioma en los casos en que el texto posea secciones en otro lenguaje, como es el caso de varios papers en español que incluyen las secciones de resumen y palabras claves en su versión en inglés.

Tal como se explicará brevemente en la sección 4.4.2, el procesamiento de *AlchemyAPI* se realiza a través de llamadas a un servidor (que para el caso del metadato idioma corresponde a access.alchemyapi.com/calls/text/ /TextGetLanguage) y los resultados se obtienen de analizar su respuesta en formato JSON. Para el metadato de idioma la clave es *language* y el valor corresponde a la cadena *spanish* para el idioma español, y *english* para el idioma inglés; a su vez, AMELOIR interpreta el resultado como *ES* y *EN* respectivamente.

Editorial y Fecha de Publicación

Para la extracción de la información editorial se tuvo en cuenta el formato como el que se muestra en el siguiente ejemplo, que figura en la mayoría de los casos de artículos de revista, libros y capítulos de libro:

"Copyright 2016, Elsevier Science (USA). All rights reserved."

Debido a que este formato puede tener muchas variaciones, se siguió el siguiente proceso para identificarlo y de esta manera extraer los metadatos editorial y fecha de publicación:

- Definición de párrafos a analizar: los *párrafos de interés* definidos para el análisis de los nombres de autor son un buen punto de partida debido a que, de estar definida una editorial, siempre se encontrará dentro de esta sección. Sin embargo, en este caso se puede refinar más el espacio de búsqueda si se parte de la premisa de que en la mayoría de los casos se presentan ciertos tokens como "Copyright" o "©". En base a esto, se decidió filtrar todos los párrafos que no posean estos tokens y analizar sintácticamente los demás.

- Aplicación de técnicas de NLP: por cada párrafo de interés encontrado anteriormente, se aplica Sentence breaking + Tokenization + POS tagging, con lo cual se obtiene como resultado un listado de tokens de la forma { "token", tag/etiqueta }. Para el caso del ejemplo, la entrada y salida se muestra en la figura 4.10.

- Mapeo con colección de tags: para simplificar el análisis, se mapean los tags a una categoría, la cual define con un nivel de abstracción mayor el rol que cumple la palabra dentro de la oración; para el caso de AMELOIR las posibles categorías son: NOUN, VERB, ADJETIVE, ADVERB, CONJ, NUMBER, PUNCT, DATE, OTHER, que corresponden a su vez a: sustantivo, verbo, adjetivo, adverbio, conjunción, número, signo de puntuación, fecha y otros. Por ejemplo, CoreNLP puede marcar sustantivos como "NN", "NNS" o "NNPS", según se trate de un sustantivo singular, sustantivo plural o nombre propio, respectivamente. En la etapa de mapeo se abstrae este tipo de palabras a la categoría "NOUN". En el ejemplo de la editorial Elsevier dado anteriormente, la salida es la siguiente:

  { "Copyright" , NOUN } { "2016" , NUMBER } { "," , PUNCT }
  { "Elsevier" , NOUN } { "Science" , NOUN } { "." , PUNCT } { "All" , OTHER }
  { "rigths" , NOUN } { "reserved" , VERB } { "." , PUNCT }

- Extracción por expresión regular: teniendo los tokens etiquetados y continuando con la premisa de que se respeta cierto patrón, se aplica una expresión regular sobre las etiquetas para determinar si la sentencia tiene el formato esperado. De ser así, se extraen los datos de interés:
  a) *Publisher*: se busca el conjunto consecutivo de sustantivos (etiqueta "NOUN"), pudiendo además incluir signos de puntuación y conjunción (etiquetas "PUNCT" y "CONJ" respectivamente), hasta encontrar la primera palabra etiquetada con "VERB" u "OTHER".
     Para el ejemplo, el resultado sería: "*Elsevier Science.*"
  b) *Published date*: se obtiene la lista de números encontrados dentro de la sentencia.
     Para el ejemplo, el resultado obtenido es: "*2016*".

En la Figura 4.10 se muestra un ejemplo completo de la entrada y resultado final del algoritmo de extracción de los metadatos editorial y fecha de publicación de un OA en AMELOIR.

La Figura 4.11 muestra el diagrama de flujo con las acciones a realizar para poder identificar los metadatos editorial y fecha de publicación de un OA en AMELOIR.

in: "Copyright 2016, Elsevier Science. All rights reserved."
out: { "Copyright" , NN } { "2016" , CD } { "," , COMMA } { "Elsevier" , NNP } { "Science" , NNP } { "." , POINT } { "All" , DT } { "rigths" , NNS } { "reserved" , VBD }
{ "." , POINT }

Fig. 4.10. Entrada y salida algoritmo de extracción metadatos Editorial y Fecha de Publicación

Formato

La forma más directa para extraer el formato del documento es a partir de la extensión del archivo; lo que se hace en este caso es validar si el nombre del archivo que representa el OA termina en .pdf, .doc o .docx, que son los formatos de archivos con los que trabaja AMELOIR, y a partir de allí se define el valor del metadato formato como PDF, DOC o DOCX según corresponda. Este es el primer metadato a obtener debido a que, a partir del mismo, se define la implementación de *DocumentService* que se utilizará para el procesamiento posterior.

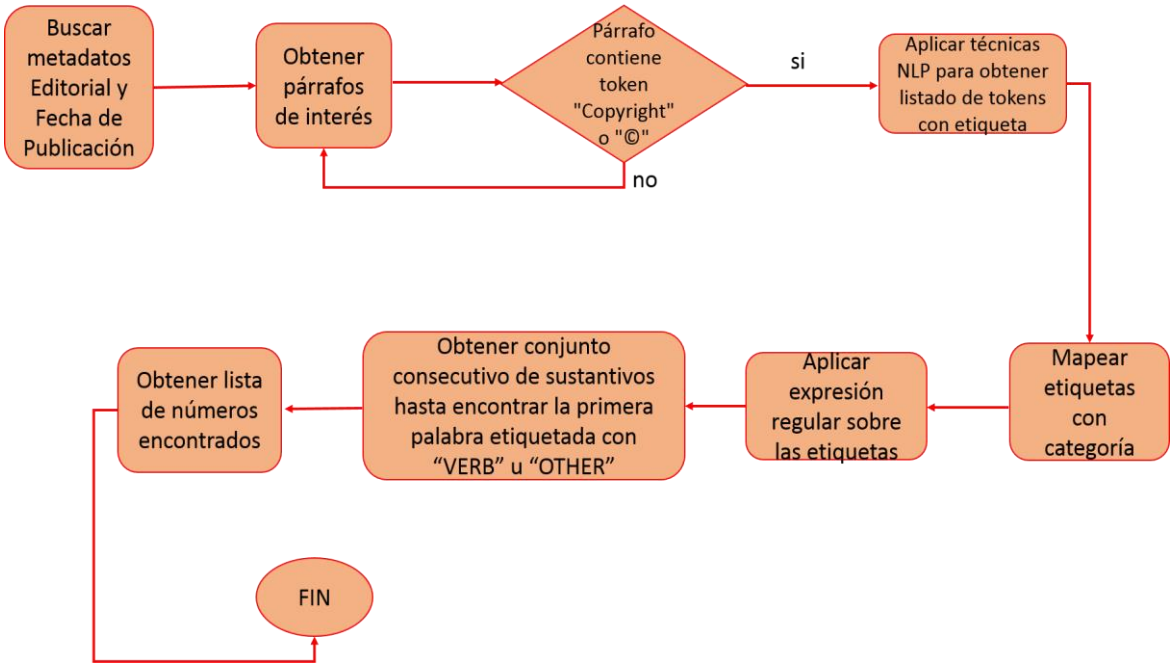

Fig. 4.11. Diagrama de flujo algoritmo de extracción metadatos Editorial y Fecha de Publicación

Identificadores

Al ser un valor estandarizado, la mejor manera de identificar este metadato es a partir de expresiones regulares. Inicialmente, se busca dentro del texto el nombre del identificador (los más comunes son "ISBN" e "ISSN"). De ser encontrado, se utilizan expresiones regulares para confirmar si lo que sigue a esta sección de texto corresponde al ISSN, si la longitud es 7 u 8 caracteres, o corresponde al ISBN, si la longitud es de 10 o 13 caracteres. La figura 4.12 muestra el diagrama de flujo que identifica las actividades llevadas a cabo para poder extraer el identificador correspondiente al OA que se está analizando.

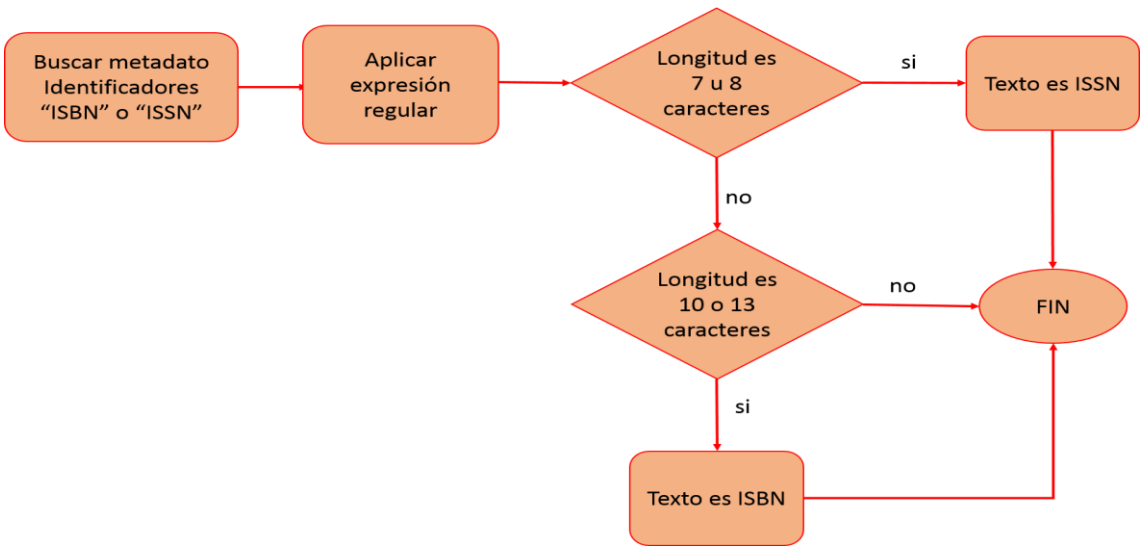

Fig. 4.12. Diagrama de flujo algoritmo de extracción metadato Identificadores

4.3.2 Metadatos Opcionales

Debido a que los siguientes metadatos fueron elegidos como opcionales en los requerimientos, no se hace el mismo tipo de análisis que los metadatos obligatorios, sino que se busca una forma simple y rápida de identificarlos. Esto implica que el extractor devolverá la mayor cantidad de metadatos posibles, sin afectar el requerimiento de tiempo de latencia.

Se decidió dar prioridad a metadatos como *Author* o *Keywords*, los cuales requieren mucho procesamiento (y por lo tanto, mayor tiempo de ejecución), de modo de obtener mejores resultados en los mismos.

Colaboradores

La extracción de los colaboradores está relacionada con la búsqueda de los autores. Una vez que se cuenta con una lista final de nombres de personas involucradas en alguna forma al documento, se analiza el lugar del texto en el que se encuentra cada uno de ellos: aquellos que se encuentran primero y más cercanos al título son los que tienen mayor posibilidad de ser autores, mientras que los colaboradores suelen tener una identificación de cuál fue su rol en el documento (editor, ilustrador, director, etc). La figura 4.13 muestra el diagrama de flujo que identifica las actividades llevadas a cabo para poder extraer los colaboradores del OA que se está analizando.

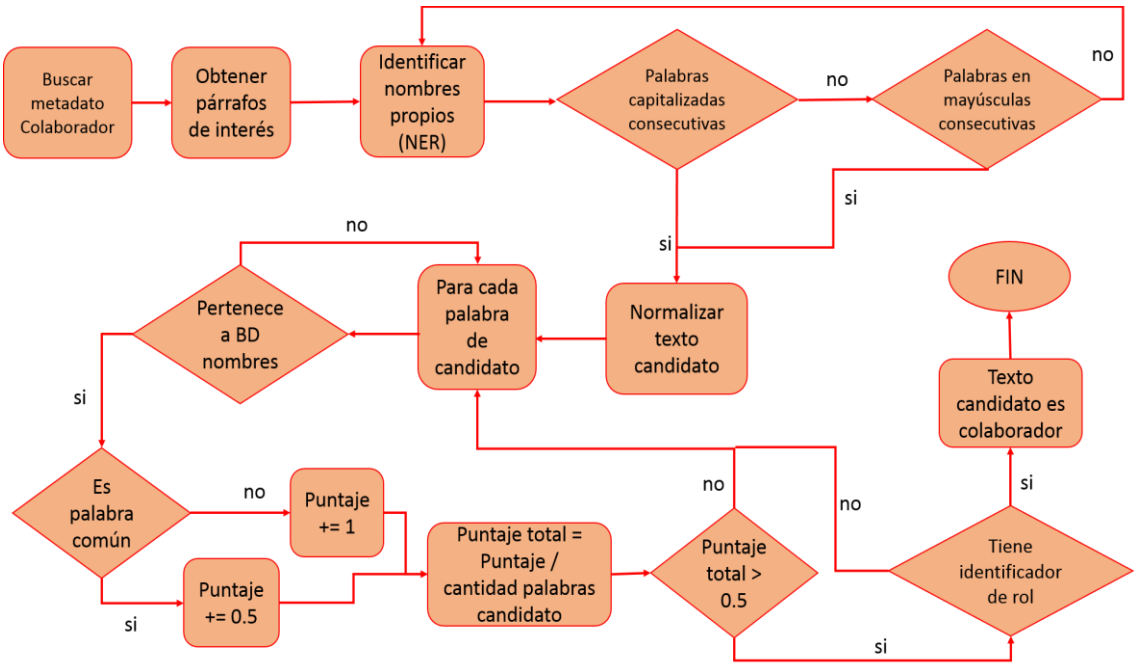

Fig. 4.13. Diagrama de flujo algoritmo de extracción metadato Colaboradores

Filiación

Se busca a partir de expresiones regulares, dentro de los mismos *párrafos de interés* que los utilizados para la extracción de nombres de autor. Al encontrar un identificador de filiación (como por ejemplo "Universidad", "Ministerio", "Centro", etc.), se analiza el texto del párrafo y se aplica NER (Named Entity Recognition) para obtener los nombres propios subsiguientes. La figura 4.14 muestra el diagrama de flujo que identifica las actividades llevadas a cabo para poder extraer la filiación correspondiente al OA que se está analizando.

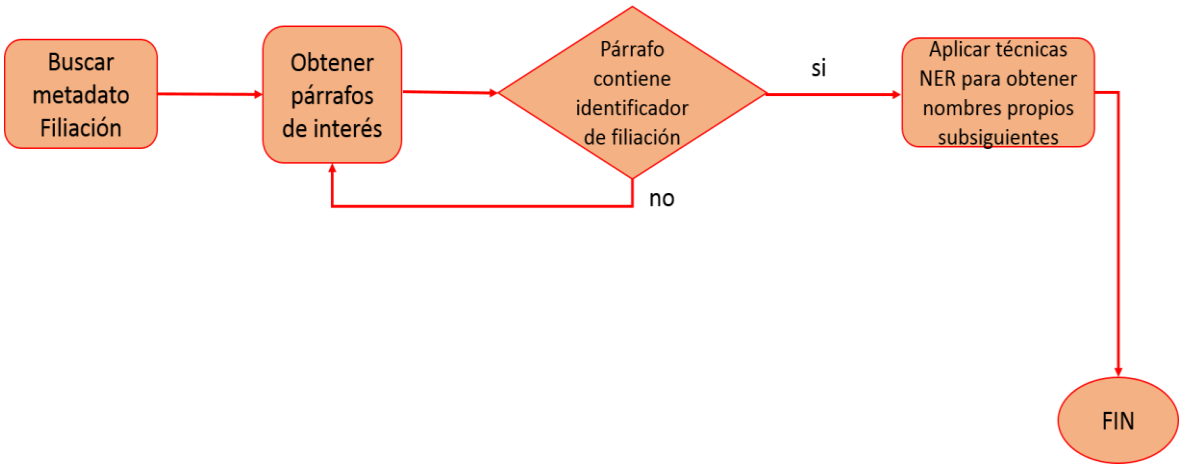

Fig. 4.14. Diagrama de flujo algoritmo de extracción metadato Filiación

Contexto

El contexto del documento se genera a partir de ciertas palabras claves presentes en el documento. Por ejemplo, el siguiente título fue extraído de un trabajo final integrador de posgrado:

"*Trabajo Final de Integración para título de Especialista en Ingeniería en Sistemas de Información*"

Al identificar palabras como "Especialista" e "Ingeniería" dentro del título, esto suma puntos como candidato al contexto POSTGRADUATE. Se analizan los mismos *párrafos de interés* definidos anteriormente, pero dando mayor peso a las palabras que se encuentren en el título. Finalmente, el contexto que mayor puntaje haya sumado, queda como definitivo (ver Figura 4.15). Los posibles valores son:

- SUPERIOR: superior
- POSTGRADUATE: posgrado

- DEGREE: grado
- SCIENCE_TECH: ciencia y tecnología
- ONG: organización no gubernamental
- CULTURAL: cultural
- BUSINESS: negocio
- OTHER: otro

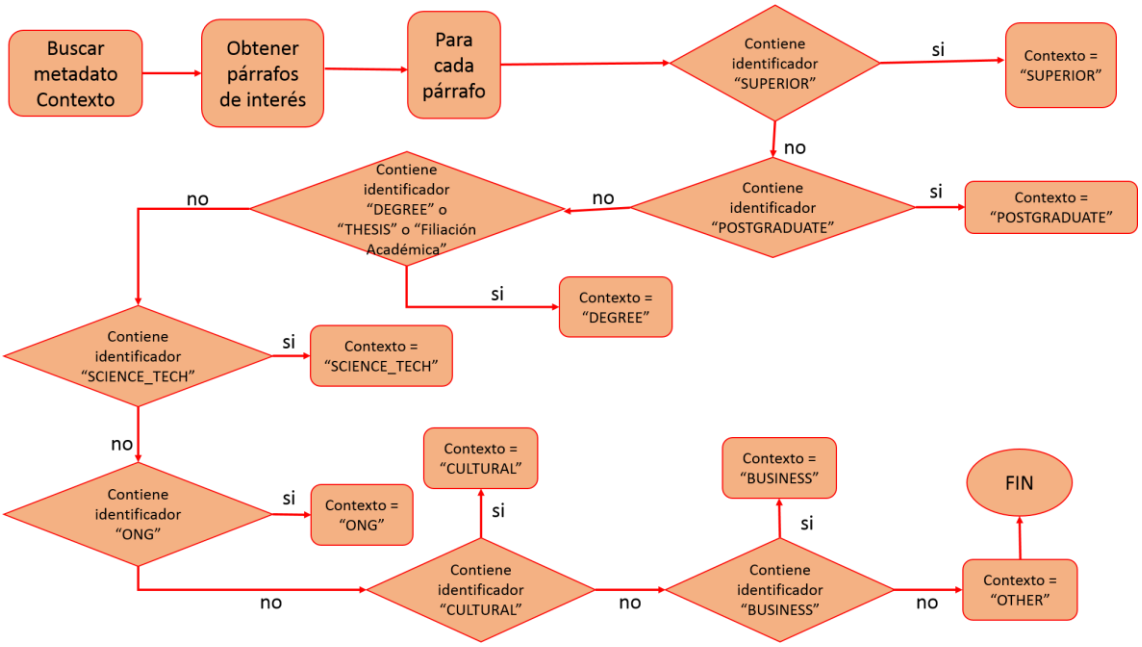

Fig. 4.15. Diagrama de flujo algoritmo de extracción metadato Contexto

Derechos

Debido a que este algoritmo de extracción será utilizado en un RI, se asume por defecto que gran parte de los documentos subidos al mismo tendrán una licencia de acceso abierto (ej. tesis, papers). También puede darse el caso que se suban documentos de acceso restringido, como por ejemplo libros o reportes técnicos. Un documento con derechos de autor restringidos tiene uno de los siguientes elementos:

- El símbolo ©, o la palabra Copyright.
- El año de la publicación.
- El nombre del propietario de los derechos.

En base a la existencia de alguno de estos identificadores en los *párrafos de interés* definidos para el análisis de los nombres de autor, se define si los derechos son Abiertos/Restringidos (OPEN/RESTRICT). La figura 4.16 muestra el diagrama de flujo que identifica las actividades llevadas a cabo para poder extraer los derechos asociados al OA que se está analizando.

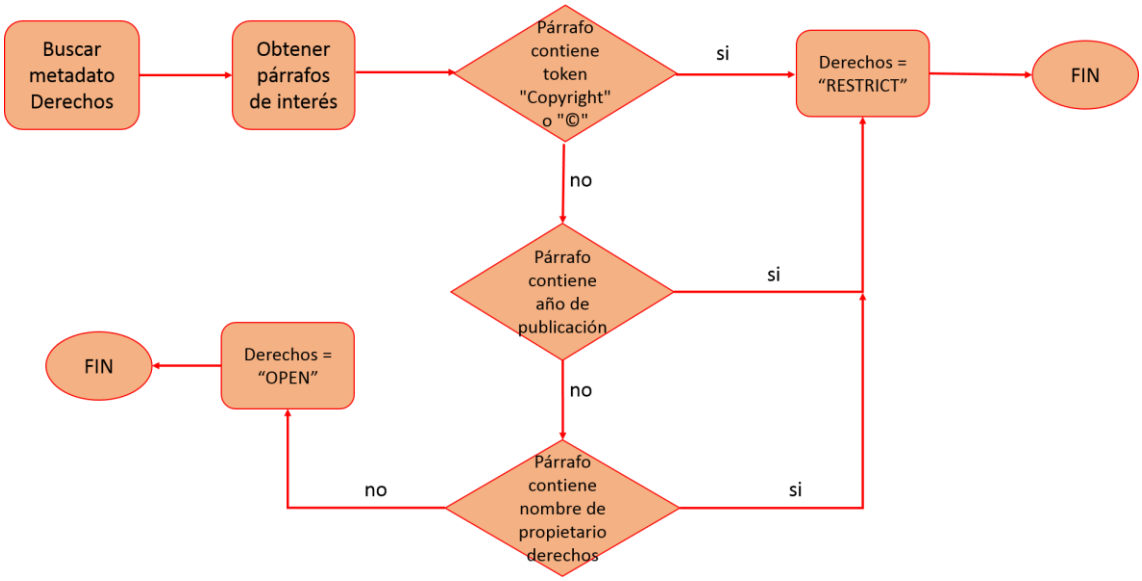

Fig. 4.16. Diagrama de flujo algoritmo de extracción metadato Derechos

Rango de Edad

Al ser el rango de edad un metadato opcional y considerando que los archivos subidos al RI se limitarán a aquellos utilizados en el contexto educativo Universitario, se definió este metadato con el valor fijo "Adultos".

Audiencia & Dificultad

Estos metadatos netamente educativos son derivados, debido a que son calculados a partir de otros:

El valor del metadato audiencia se define a partir de la combinación de los metadatos rango de edad y contexto: *Target audience* = age range + context, pudiendo tener los valores:

- EXPERT: experto
- KNOWLEDGE: conocimiento
- NO_KNOWLEDGE: sin conocimiento

Teniendo en cuenta que para el caso de AMELOIR rango de edad se definió como valor fijo "Adultos", la audiencia objetivo se deduce a partir del metadato contexto como se muestra en la Tabla 4.3.

Tabla 4.3. Definición del metadato audiencia a partir del contexto.

Contexto	Audiencia
POSTGRADUATE	EXPERT
SCIENCE_TECH	EXPERT
SUPERIOR	KNOWLEDGE
DEGREE	KNOWLEDGE
BUSINESS	KNOWLEDGE
ONG	NO_KNOWLEDGE
CULTURAL	NO_KNOWLEDGE
OTHER	NO_KNOWLEDGE

Para el caso del metadato dificultad, su valor se deriva de la combinación de los metadatos audiencia y contexto: *Difficulty* = audience + context, pudiendo tener los valores:

- SIMPLE: simple
- MEDIUM: medio
- DIFFICULT: difícil

La Tabla 4.4 muestra los valores del metadato dificultad de acuerdo al metadato contexto y audiencia usada por AMELOIR para extraer este metadato.

Tabla 4.4. Definición del metadato dificultad a partir del contexto y la audiencia.

Contexto	Audiencia	Dificultad
POSTGRADUATE	EXPERT	DIFFICULT
SCIENCE_TECH	EXPERT	DIFFICULT
SUPERIOR	KNOWLEDGE	MEDIUM
DEGREE	KNOWLEDGE	MEDIUM
BUSINESS	KNOWLEDGE	MEDIUM
ONG	NO_KNOWLEDGE	SIMPLE
CULTURAL	NO_KNOWLEDGE	SIMPLE
OTHER	NO_KNOWLEDGE	SIMPLE

4.4 Tecnologías Utilizadas

Toda herramienta evaluada, tanto para ser usada en el desarrollo de la solución, como para ser parte de la misma, debía cumplir con el requisito no funcional de ser gratuita y de código abierto, ya que el proyecto tiene como destino principal el entorno del RI de una facultad perteneciente a una Universidad pública.

Además del requisito anterior, para cada librería utilizada en el proyecto se tuvo en cuenta lo siguiente:

- *Capacidad de integración con el proyecto*: para poder importar una librería al proyecto ésta debía tener una interfaz compatible con Java (Oracle, 1995). Se priorizaron aquellas alojadas en el repositorio Maven (Zyl, 2002), a través del cual fueron centralizadas las dependencias del proyecto.

- *Facilidad de uso*: sin una documentación adecuada, una librería puede ser muy complicada de usar, dependiendo de la implementación de sus interfaces. Es por esto que se priorizaron las librerías bien documentadas, o cuyas interfaces resultaban sencillas de entender.

- *Calidad de los resultados*: este criterio fue el de mayor peso al momento de decidir sobre la utilización, o no, de una librería en particular. Las características evaluadas para los resultados fueron: la capacidad de procesarlos mediante código, y qué tanto se ajustaban a las técnicas de tratamiento de lenguaje natural, según correspondía al tipo de librería.

- *Tiempos de respuesta*: para los casos de procesamientos más complejos efectuados por el algoritmo de extracción AMELOIR, se evaluó que el tiempo demandado, antes de dar un resultado final, no supere los 60 segundos definidos en los requisitos no funcionales.

Algunas herramientas empleadas presentan un modelo de Software como un Servicio (SaaS, de sus siglas en inglés para Software as a Service). Este es un modelo de distribución de software donde el soporte lógico y los datos que maneja se alojan en servidores de una compañía de tecnologías de información y comunicación (TIC), a los que se accede vía Internet desde un cliente. Debido a su naturaleza, además de los criterios descritos anteriormente, para estas herramientas se debió tener en cuenta su disponibilidad. Si los servidores no eran consistentes en su capacidad de brindar respuestas, se decidió que no sean incluidas en la implementación del algoritmo de extracción.

Para la elección de las herramientas usadas en la gestión del proyecto, como el entorno de desarrollo, gestión de dependencias, gestión de versiones y el mismo lenguaje de

programación, fueron decisivos los conocimientos y experiencia previa de los participantes en el proceso de desarrollo del algoritmo AMELOIR.

4.4.1 APIs y Librerías

A lo largo de la implementación fue necesario recurrir a distintas librerías Java y APIs para lograr acceder y manipular los datos almacenados en los archivos analizados por el algoritmo de extracción, y así poder obtener la información deseada. A continuación se listan las librerías y APIs que resultaron útiles y que fueron usadas, luego de investigar las alternativas disponibles.

1. *Extracción de texto*

Para la extracción y análisis de texto fueron empleadas varias APIs. Si bien algunas de las herramientas que ofrecen son redundantes, se eligió una u otra dependiendo de la calidad de los resultados de cada una, y la facilidad de uso de las mismas.

a) Apache POI (Apache, Apache POI - the Java API for Microsoft Documents, 2002)

Provee APIs Java para manipular varios formatos de archivos basados en los estándares Office Open XML (OOXML) y el formato de Documento Compuesto OLE 2, de Microsoft (OLE2). Con estas APIs se puede leer y escribir en archivos MS Excel, MS Word y MS PowerPoint usando código Java.

En particular, se usaron dos de sus componentes: HWPF y XWPF para archivos .doc y .docx, respectivamente. Resultaron de uso sencillo, relativamente, permitiendo acceder a la estructura de texto de los archivos sin mayores inconvenientes.

b) Snowtide PDFxStream (Snowtide, 2001)

Permite extraer información de cualquier archivo PDF, sea texto, imágenes o formularios interactivos. Posee una API diseñada específicamente para Java, por lo que no se presentaron problemas para integrarla al proyecto.

Teniendo en cuenta que la especificación para archivos PDF es muy compleja en sí misma, el manejo de este formato de archivos resultó ser complejo. Si bien la librería permite extraer toda la información, la calidad de la misma depende de cómo fue creado el documento en primera instancia, ya que la especificación PDF es una recomendación de cómo debería ser la estructura de un archivo PDF.

2. *Procesamiento de lenguaje natural*

a) AlchemyAPI (IBM, AlchemyAPI, 2009)

AlchemyAPI es una compañía que, entre otras cosas, emplea aprendizaje computarizado para realizar procesamiento de lenguaje natural a partir de texto, tal como se introdujo en la sección 2.5. AlchemyAPI ofrece su tecnología en la forma de SaaS, lo cual implica que el procesamiento se solicita a través de llamadas a un servidor y los resultados se obtienen de analizar su respuesta. Entre las funciones que brinda resultaron interesantes dos: extracción de entidades y detección de lenguaje.

Para utilizar este servicio fue necesario obtener una clave (API Key) en el sitio de la empresa, que luego debe insertarse en la llamada al servidor. Este proceso es gratuito, pero para no incurrir en costos se debió elegir el plan gratuito, que limita el número de sucesos (respuestas del servidor) a 1000 por día.

b) Stanford CoreNLP (Manning, y otros, 2014)

Esta API ofrece un conjunto de herramientas de análisis de lenguaje natural. Su objetivo es facilitar la aplicación de varios análisis lingüísticos a un fragmento de texto. Sus herramientas pueden correrse consecutivamente en un pipeline con escasas líneas de código, de forma flexible y extensible.

Stanford CoreNLP integra muchas de las herramientas para NLP: POS tagging, NER, sistema de resolución de correferencias, entre otras. Provee los bloques básicos para construir aplicaciones de análisis de texto de alto nivel y específicas de un dominio. Con simples opciones puede elegirse qué herramientas deben activarse durante la ejecución, de las cuales para la implementación de AMELOIR se utilizó la de POS tagging.

c) Apache Lucene Core (Apache, Apache Lucene Core, 2011)

Es una librería de recuperación de información gratuita y de código abierto. Si bien su desempeño y características se ven mejor aprovechados para motores de búsqueda, resultó ser de gran utilidad para reconocer las denominadas Stop Words o palabras vacías, mencionadas en la sección 2.3, y así poder evitar su análisis.

d) Freeling (Padró & Stanilovsky, 2012)

Está diseñada para ser usada como librería externa que provee funcionalidades de análisis de lenguaje (análisis morfológico, NER, POS tagging, análisis sintáctico, etc.) para diversos idiomas. Mediante esta librería, en combinación con Apache Lucene, se realizó el proceso de POS tagging para obtener las palabras más relevantes y, a partir de ellas, detectar las palabras clave o keywords.

Al tratarse de una librería externa, no se encuentra en el repositorio Maven a través del cual se centralizaron todas las dependencias del proyecto. Está escrita originalmente en C++, por lo que no es nativamente compatible con el código del proyecto.

Para poder hacer uso de sus funciones desde un programa Java, debe compilarse una API en un archivo JAR, e instalar librerías DLL accesibles desde variables de entorno del sistema operativo. Para el desarrollo se utilizó esta librería en su versión 3.1 (2014), siendo la más reciente la versión 4.0 (2016). Esto se debió a que la versión 3.1 se encuentra disponible para descargar y utilizar en Java, mientras que la última versión debe ser recompilada, un proceso en extremo difícil y que no fue posible concretar, ya que no existe documentación específica para realizarlo en plataformas Windows.

4.4.2 Repositorio Institucional

En lo que respecta al RI para el cual fue diseñado el algoritmo de extracción de metadatos AMELOIR, se presentan a continuación algunos de los elementos tecnológicos involucrados en el proceso de desarrollo del algoritmo.

1. DSpace (DuraSpace, 2002)

Es un paquete de software de código abierto para repositorios, típicamente empleado para crear repositorios de Acceso Abierto de instituciones educativas y/o publicaciones de contenido digital. Aunque DSpace comparte algunas características con sistemas de gestión de contenido o de gestión de documentos, el software de repositorio de DSpace cubre una necesidad específica como sistema de archivos digitales, focalizado en el almacenamiento a largo plazo, el acceso y la preservación de contenidos digitales.

DSpace es un conjunto cooperativo de aplicaciones web Java y programas utilitarios que mantienen un almacén de activos digitales, y sus metadatos asociados. Las aplicaciones web proveen interfaces para la administración, depósito, ingreso, búsqueda y acceso. El almacén de activos es mantenido en un sistema de archivos, mientras que los metadatos, incluyendo la

información de acceso y configuración, se almacenan en una base de datos relacional, con soporte de bases de datos PostgreSQL y Oracle. Actualmente DSpace soporta dos interfaces web: JSPUI y XMLUI (también conocida como Manakin) basada en Apache Cocoon, que usa XML y XSLT (Extensible Stylesheet Language Transformations).

2. XStream (Walnes, 2004)

Es una librería Java para serializar objetos a XML, y viceversa. XStream usa reflexión para descubrir la estructura de objetos a serializar en tiempo de ejecución, sin requerir modificarlos. Puede serializar campos internos, incluyendo los private y final, y también las clases no públicas e internas.

Fue necesario recurrir al uso de esta librería para poder llevar a cabo el flujo de información entre el proceso de envío de objetos a DSpace. Como se mencionó anteriormente, las estructuras y flujos en DSpace resultaron muy complicadas de modificar, por lo que se mitigó esta dificultad escribiendo y leyendo desde el disco la información extraída antes de guardarla en el repositorio.

4.4.3 Otros Diseños Considerados

El diseño de AMELOIR se vio enmarcado por diferentes alternativas que fueron evolucionando conforme se revisaban las necesidades de los usuarios y la viabilidad de su implementación; a continuación se da un vistazo a dichas consideraciones de diseño.

En el d*iseño inicial* consideraba la implementación de un proxy de adaptadores, dependiendo el formato del OA que se estaba procesando, para convertir todo el contenido del mismo a formato TXT; incluía la utilización de ParsCit y la verificación de datos contra diccionarios de vocabulario controlado (tesauros), que buscaba mejorar la precisión de la extracción de metadatos determinando el dominio del documento analizado para eliminar ambigüedades. Así mismo, se proponía hacer uso únicamente de Alchemy para extraer idioma y palabras clave. Sin embargo, se encontró lo siguiente:

- *Proxy de adaptadores*: el objetivo del proxy era tomar el OA en su formato original y convertir todo el contenido del mismo a formato TXT sin realizar ningún tipo de análisis. Sin embargo, en vez de pasar directamente todo el documento a un archivo TXT, lo que se hizo fue analizarlo y definir las secciones del documento a grandes rasgos: identificando por un lado los encabezados y pies de página (en donde suele haber información clave), por otro lado,

identificando el título del documento y finalmente todo el texto "útil"; es decir, pasar a texto plano todo eliminando cosas como los párrafos de las secciones *Agradecimientos, Anexos* y *Referencias*, ya que no aportan información significativa para extraer metadatos.

Esto conllevó a descartar también el uso de *plantillas* para la extracción de metadatos conforme a la colección seleccionada para almacenar el OA, dado que, en general, el diseño de AMELOIR consideraba analizar o descartar las secciones del documento anteriormente mencionadas.

- *ParsCit*: tal como se mencionó en la sección 3.5, esta herramienta puede ser utilizada para extraer la estructura lógica de documentos científicos e inicialmente se consideró también para la extracción de información de las referencias del documento. Sin embargo, debido a su baja disponibilidad (el servicio responde casi constantemente que se encuentra caído) no se tuvo en cuenta dentro del proceso de desarrollo. Por otra parte, se decidió omitir las referencias al momento de extraer metadatos debido a que las mismas no aportan información útil para los metadatos requeridos.

- *Uso de tesauros*: se consideró realizar la verificación de metadatos extraídos contra diccionarios de vocabulario controlado, dado que los documentos que potencialmente pueden ser enviados al RI del CIDISI, para el cual fue diseñado AMELOIR, se pueden encasillar en un dominio en particular, y se podrían establecer términos normalizados para los temas más comunes tratados en este entorno. A su vez, los términos normalizados pueden ser relacionados con los equivalentes en otros idiomas, con lo cual se mejorarían en gran medida los posteriores procesos de búsquedas en el RI, enlazando documentos por temas, en múltiples idiomas. Sin embargo, se presentaron dificultades a la hora de acoplar dichas herramientas en el desarrollo de AMELOIR, dado que las herramientas analizadas están enfocadas únicamente en el idioma inglés, porque los resultados obtenidos resultaron de difícil procesamiento en conjunto con el algoritmo propuesto, y por otro lado, no se ofrecía la posibilidad de descargar una base de datos para efectuar consultas locales.

- *Alchemy*: teniendo en cuenta que según el análisis que se hizo para varios documentos Alchemy comete muchos errores (sobre todo para textos en español), se consideró tomar estos resultados como parciales y hacerlos más confiables a partir de la utilización en conjunto de otras herramientas, como las técnicas de procesamiento de lenguaje natural NLP.

Es así como el diseño de AMELOIR fue evolucionando hasta llegar a ser lo que se presentó en la sección 4.1, que fue evaluado en los casos de estudio que se presentan a continuación.

4.5 Casos de Estudio

Con el fin de evaluar el funcionamiento y comportamiento del algoritmo para la extracción de metadatos AMELOIR en el entorno DSpace, se seleccionaron tres tipos de objetos de aprendizaje:

1. Artículo de revista.

2. Libro.

3. Tesis de posgrado.

A continuación se mostrarán los resultados obtenidos para cada uno de dichos documentos, donde puede observarse que los tiempos de procesamiento de extracción de metadatos se mantuvieron menores a los 60 segundos, especificados en los requisitos no funcionales que se presentaron en la sección 4.1.

4.5.1 Artículo de Revista

Para este caso de estudio se seleccionó el artículo denominado “FAIA: Framework para la enseñanza de agentes en IA” publicado en la Revista Iberoamericana de Informática Educativa en el año 2008, sus autores son Jorge Roa, Milagros Gutiérrez y Georgina Stegmayer (Roa, Gutiérrez, & Stegmayer, 2008). El documento se sube a DSpace en formato DOCX, y al ser procesado por AMELOIR genera los resultados que se muestran en la Tabla 4.5.

El tiempo total de procesamiento de AMELOIR fue de 8 segundos, durante los cuales se llevaron a cabo los siguientes pasos para la extracción de metadatos:

- *AMXService* solicita a *DocumentServiceFactory* el servicio necesario para manejar el OA, que en este caso corresponderá a una instancia de la clase *DOCXService*, teniendo en cuenta que el documento se encuentra en formato DOCX, que corresponderá al valor del metadato *format*. Dicha instancia será una representación estructurada del documento, con toda la información extraída del mismo.

- En el siguiente paso se utilizó el servicio *NLPService* para extraer el metadato *identifier*; en este caso el documento contiene el identificador "ISSN: 16994574", que se encuentra de forma directa en el texto.

- El identificador extraído se utilizó para hacer una consulta a *BooksService* e intentar obtener los metadatos desde *Worldcat*, teniendo en cuenta que el tipo de recurso es un *Article*.

- La consulta en *Worldcat* no retorna información sobre el recurso, por lo que se procede a extraer los metadatos *author* y *title* para intentar hacer una nueva consulta sobre estos valores.

- El metadato *title*, se obtuvo utilizando el servicio *DocumentService*, donde se analizan los párrafos de interés de las 3 primeras páginas del documento, identificándolo debido a que cuenta con un mayor tamaño de fuente.

- Para el caso de *author*, se obtuvo utilizando el servicio *NLPService*, donde se analizan los párrafos de interés de las 3 primeras páginas del documento, identificándolos debido a que son los únicos nombres propios que se encuentran en esta sección del documento.

- Con los metadatos *title* y *author* se realiza una nueva consulta desde *Worldcat*. Esta vez se obtuvieron algunos metadatos adicionales como *Publisher* y *Publisher Date*.

- Los demás metadatos como *description, keywords, filiation, context, rights, age range* (con valor fijo "ADULTS"), *audience* (derivado de age range + context)*, language* (haciendo uso de AlchemyAPI), se obtuvieron aplicando los algoritmos descritos en las secciones 4.3.1 y 4.3.2.

En las figuras 4.17 a 4.19 se presentan los pasos llevados a cabo para cargar el objeto de aprendizaje en DSpace, conforme a lo especificado en los requisitos funcionales de la sección 4.1 de este documento.

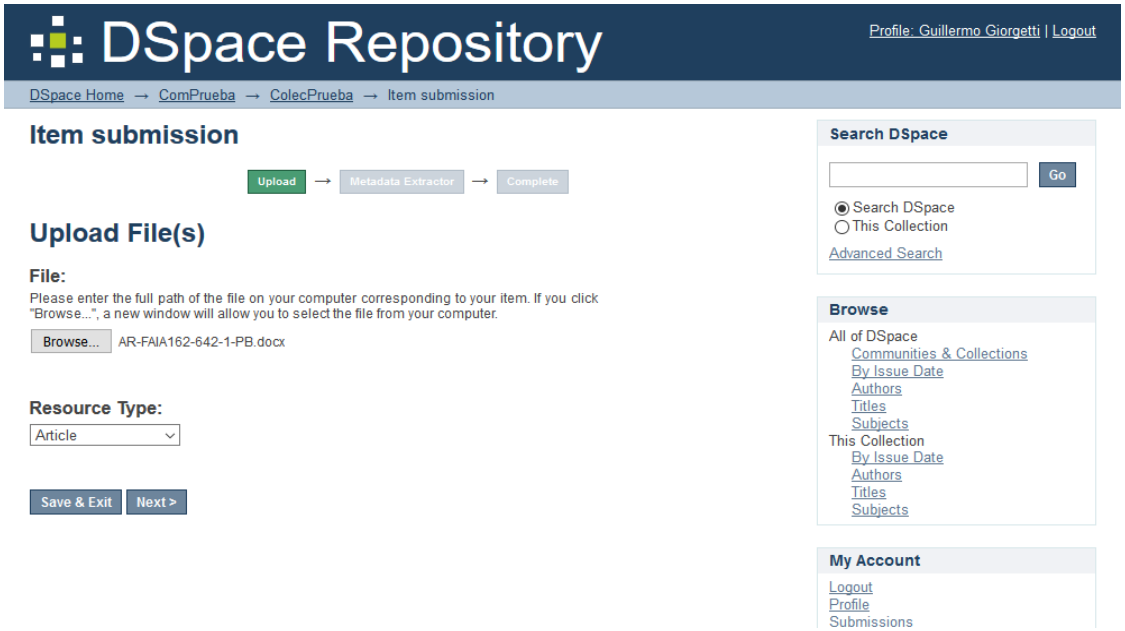

Fig. 4.17 Carga del archivo correspondiente al objeto de aprendizaje

Para comenzar el envío, haciendo uso del botón "*Browse…*" marcado en la Figura 4.17, se ubica el archivo correspondiente al objeto de aprendizaje en la máquina local y luego se selecciona el tipo de recurso al que pertenece el documento en el combo box *Resource Type*, donde fueron incluidos los tipos de recurso o colecciones definidas en la sección 4.2 de este documento: artículo, libro, capítulo de libro, paper, tesis, paper y reporte técnico.

Luego del procesamiento, se muestra la pantalla de visualización de resultados del proceso de extracción de metadatos, correspondiente a la Figura 4.18.

Una vez que se completa el envío haciendo uso del botón *Complete submission*, los datos son almacenados junto al objeto de aprendizaje subido. Los resultados que se muestran en el último paso dependen de cuáles fueron los metadatos hallados, por lo que dado el caso que se hallen menos o ninguno de los definidos como obligatorios en la sección 4.3.1, o para tener la opción de modificarlos, en la página de envío completo se provee un enlace directo a la sección de edición de información (*Go to the "Edit item" section*), tal como se muestra en la Figura 4.19.

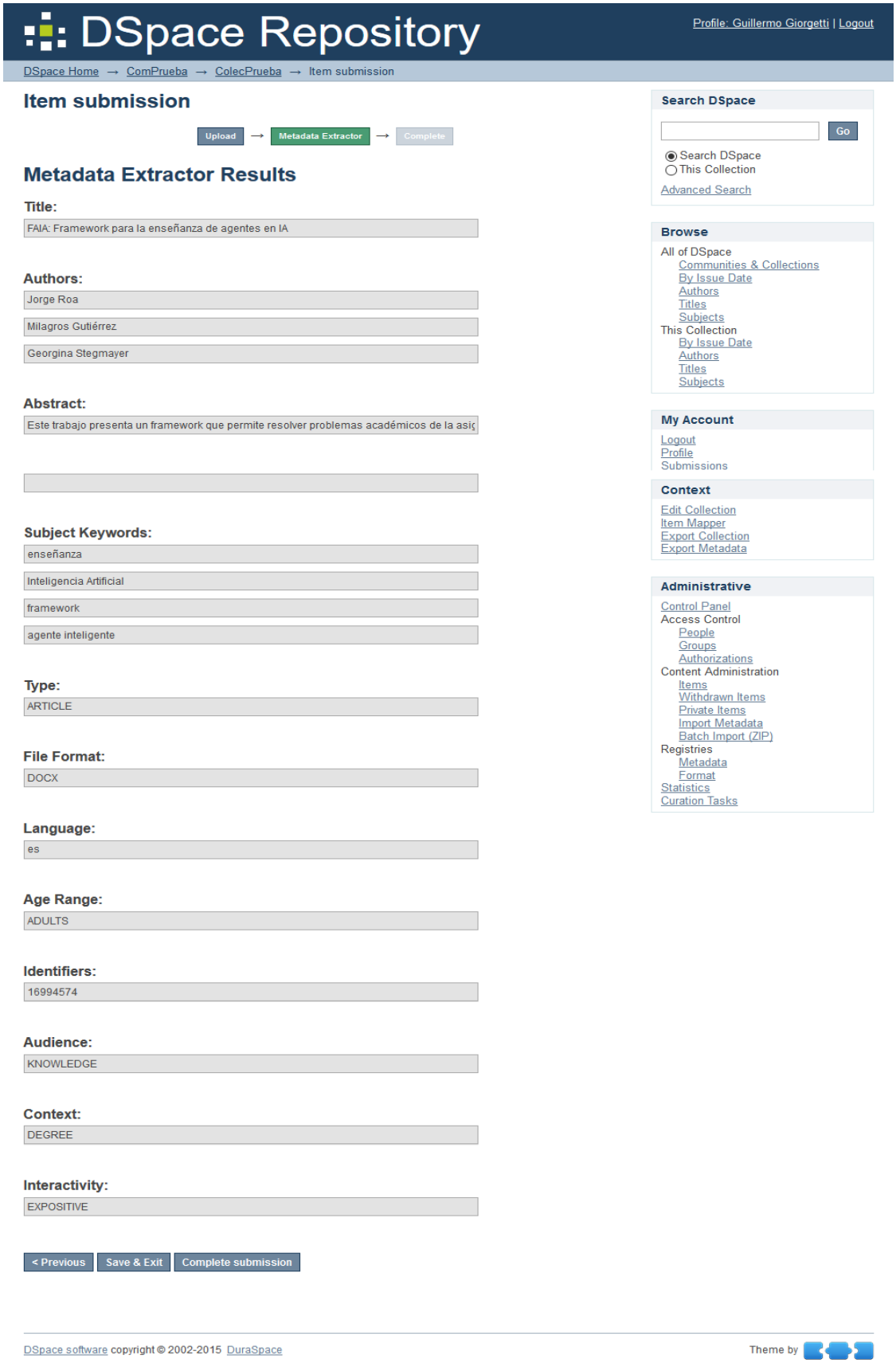

Fig. 4.18. Resultados AMELOIR para artículo de revista

Para los siguientes casos de estudio solo se presentarán las figuras correspondientes a los resultados del extractor de metadatos AMELOIR integrado con DSpace, dado que el proceso de carga es el mismo en todos los casos.

Tabla 4.5. Resultados para un artículo de revista.

Metadato	Valor
Title	Framework para la enseñanza de agentes en IA
Authors	JORGE ROA MILAGROS GUTIERREZ GEORGINA STEGMAYER
Description	Este trabajo presenta un framework que permite resolver problemas académicos de la asignatura Inteligencia Artificial (IA) aplicando agentes basados en objetivos que usan búsqueda para la toma de decisión. Con el uso del framework, denominado FAIA: Framework para un Agente que resuelve problemas de IA, el alumno puede centrar su atención en la definición de las propiedades del agente y su estrategia para decidir qué acción emprender en cada interacción con el ambiente, sin tener que preocuparse por el simulador del ambiente en el cual se desenvuelve el agente. Para su desarrollo se usaron técnicas de diseño orientado a objeto y patrones de diseño. El objetivo de este trabajo es proveer una herramienta flexible con la cual el alumno pueda aprender a construir agentes inteligentes, comprendiendo la relación de éste con su ambiente. De esta forma pueden proponerse problemas a resolver más complejos, requiriendo menos tiempo en la implementación de una solución.
Keywords	enseñanza Inteligencia Artificial framework agente inteligente
Publisher	© ADIE Asociación
Published Date	Enero Diciembre 2008
Filiation	Universidad Tecnológica Nacional Fac. Reg. Sta. Fe Lavaise 610, 3000 Santa Fe
Resource Type	ARTICLE
Format	DOCX
Languages	ES
Age Range	ADULTS
Industry Identifier	ISSN: 16994574
Audience	KNOWLEDGE
Context	DEGREE
Interactivity	EXPOSITIVE
Rigths	OPEN

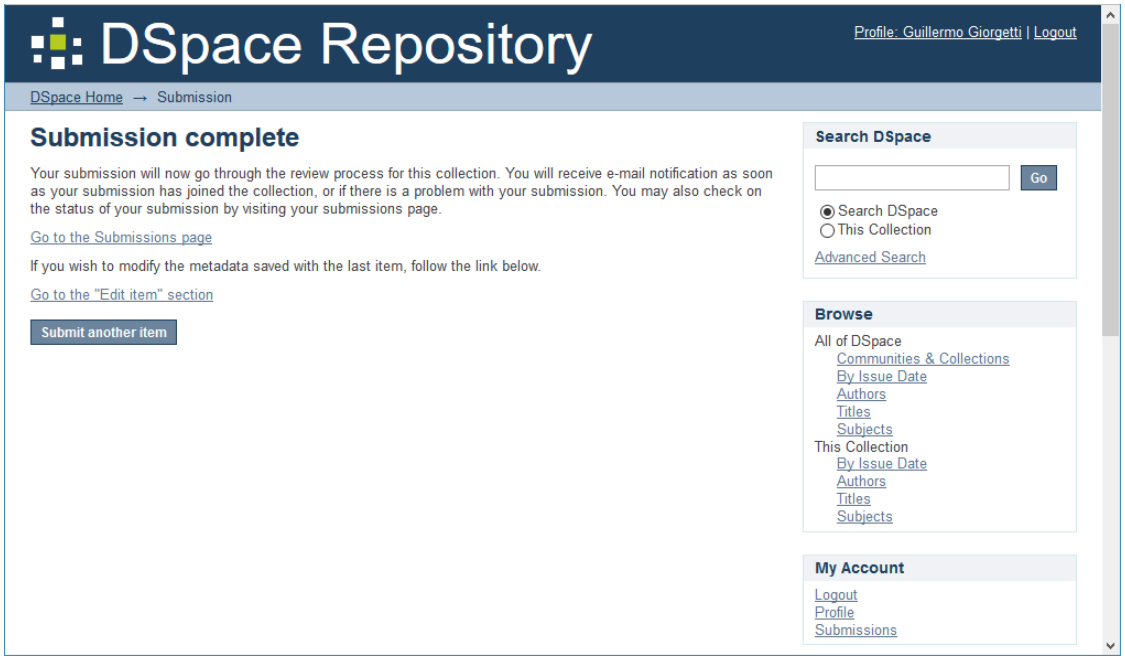

Fig. 4.19. Finalización de la carga del archivo correspondiente al objeto de aprendizaje

4.5.2 Libro

Para este caso de estudio se seleccionó el libro con título "Computer Networks" que fue publicado en 2011 y cuyos autores son Andrew S. Tanenbaum y David Wetherall (Tanenbaum & Wetherall, 2011). El documento se sube a DSpace en formato PDF y, al ser procesado por AMELOIR, produce los resultados presentados en la Tabla 4.6.

El tiempo total de procesamiento de AMELOIR fue de 48 segundos, durante los cuales se llevaron a cabo los siguientes pasos para la extracción de metadatos:

- *AMXService* solicita a *DocumentServiceFactory* el servicio necesario para manejar el OA, que en este caso corresponderá a una instancia de la clase *PDFService*, teniendo en cuenta que el documento se encuentra en formato PDF, que corresponderá al valor del metadato *format*. Dicha instancia será una representación estructurada del documento, con toda la información extraída del mismo.

- En el siguiente paso se utilizó el servicio *NLPService* para extraer el metadato *identifier*; en este caso el documento contiene dos identificadores: "ISBN: 0132126958, ISBN: 9780132126953", que se encuentran de forma directa en el texto. Debido a que este metadato se clasificó como multivaluado es permitido que tenga más de una valor.

- El identificador extraído se utilizó para hacer una consulta a *BooksService* e intentar obtener los metadatos desde *Google Books*, teniendo en cuenta que el tipo de recurso es un *Book*.

- La consulta en *Google Books* retorna los siguientes metadatos: *authors*, *title*, *description*, p*ublisher, publisher date*.

- Los demás metadatos como *contributors, keywords, filiation, context, rights, age range* (con valor fijo "ADULTS"), *audience* (derivado de age range + context)*, language* (haciendo uso de AlchemyAPI), se obtuvieron aplicando los algoritmos descritos en las secciones 4.3.1 y 4.3.2.

Tabla 4.6. Resultados para un libro.

Metadato	Valor
Title	Computer Networks
Authors	Andrew S. Tanenbaum David Wetherall
Contributors	ALAN FISCHER ALLISON MICHAEL ARON MARVIN BARBARA DANIEL DANIEL SANDIN JASON CONSALVO JEFF HOLCOMB JOE RUDDICK KATHRYN FERRANTI LEHIGH PHOENIX LINDA KNOWLES MARCIA HORTON MELINDA HAGGERTY MICHAEL HIRSCH PATRICE JONES RACHEL HEAD SUSAN PARADISE TANENBAUM ANDREW TRACY DUNKELBERGER VINCE O'BRIEN
Description	Computer Networks, 5/e is appropriate for Computer Networking or Introduction to Networking courses at both the undergraduate and graduate level in Computer Science, Electrical Engineering, CIS, MIS, and Business Departments. Tanenbaum takes a structured approach to explaining how networks work from the inside out. He starts with an explanation of the physical layer of networking, computer hardware and transmission systems; then Works his way up to network applications. Tanenbaum's in-depth application coverage includes email; the domain name system; the World Wide Web (both client- and server-side); and multimedia (including voice over IP, Internet radio video on demand, video conferencing, and streaming media. Each chapter follows a consistent approach: Tanenbaum presents key principles, then illustrates them utilizing real-world example networks that run through the entire book—the Internet, and wireless

	networks, including Wireless LANs, broadband wireless and Bluetooth. The Fifth Edition includes a chapter devoted exclusively to network security. The textbook is supplemented by a Solutions Manual, as well as a Website containing PowerPoint slides, art in various forms, and other tools for instruction, including a protocol simulator whereby students can develop and test their own network protocols.
Metadato	**Valor**
Keywords	network computer protocol layer system internet service model
Publisher	Prentice Hall
Published Date	2011
Filiation	University of Washington Seattle, WA
Resource Type	BOOK
Format	PDF
Languages	EN
Age Range	ADULTS
Industry Identifier	ISBN: 0132126958 ISBN: 9780132126953
Audience	KNOWLEDGE
Context	DEGREE
Interactivity	EXPOSITIVE
Rigths	RESTRICT

4.5.3 Tesis de Posgrado

La tesis tomada como caso de estudio tiene por título “Un Framework basado en las Tecnologías de la Web Semántica para dar soporte a la Generación de Especificaciones de Requerimientos de Software”, elaborada por Verónica Castañeda para su tesis de Maestría en Ingeniería en Sistemas de Información. El archivo está en formato DOC, y cuando es procesado por AMELOIR produce la salida que se muestra en

El tiempo total de procesamiento de AMELOIR fue de 41 segundos, durante los cuales se llevaron a cabo los siguientes pasos para la extracción de metadatos:

- *AMXService* solicita a *DocumentServiceFactory* el servicio necesario para manejar el OA, que en este caso corresponderá a una instancia de la clase *DOCService*, teniendo en cuenta que el documento se encuentra en formato DOC, que corresponderá al valor del metadato *format*. Dicha instancia será una representación estructurada del documento, con toda la información extraída del mismo.

- Como en este caso se trata de una tesis de posgrado que no cuenta con el metadato *identifier*, se procede de inmediato a extraer, en primera instancia, los metadatos *author* y *title*.

- El metadato *title*, se obtuvo utilizando el servicio *DocumentService*, donde se analizan los párrafos de interés de las 3 primeras páginas del documento, identificándolo debido a que cuenta con un mayor tamaño de fuente.

- Para el caso de *author*, se obtuvo utilizando el servicio *NLPService*, donde se analizan los párrafos de interés de las 3 primeras páginas del documento, identificándolos debido a que son nombres propios que se encuentran se encuentran primero y más cercanos al título.

- Junto con el metadato *author*, fue posible extraer el valor del metadato *contributor*, dado que también se trataba de nombres propios, pero acompañados de una identificación de cuál fue su rol, que en este caso corresponde a "Jurado de Tesis".

- Los demás metadatos como *keywords, filiation, published date, context, interactivity, age range* (con valor fijo "ADULTS"), *audience* (derivado de age range + context), *language* (haciendo uso de AlchemyAPI), se obtuvieron aplicando los algoritmos descritos en las secciones 4.3.1 y 4.3.2.

Tabla 4.7. Resultados para un tesis de posgrado

Metadato	Valor
Title	Un Framework basado en las Tecnologías de la Web Semántica para dar soporte a la Generación de Especificaciones de Requerimientos de Software
Authors	LUCIANA BALLEJOS MARÍA LAURA CALIUSCO VERÓNICA CASTAÑEDA
Contributors	GRACIELA HADAD MARIELA RICO SILVIO GONNET
Keywords	requerimiento especificación generación ers información actividad dominio conocimiento
Published Date	Noviembre 2012
Filiation	Universidad Tecnológica Nacional Facultad Regional Santa Fe
Resource Type	THESIS
Format	DOC

Metadato	Valor
Languages	ES
Age Range	ADULTS
Audience	EXPERT
Context	POSTGRADUATE
Interactivity	EXPOSITIVE

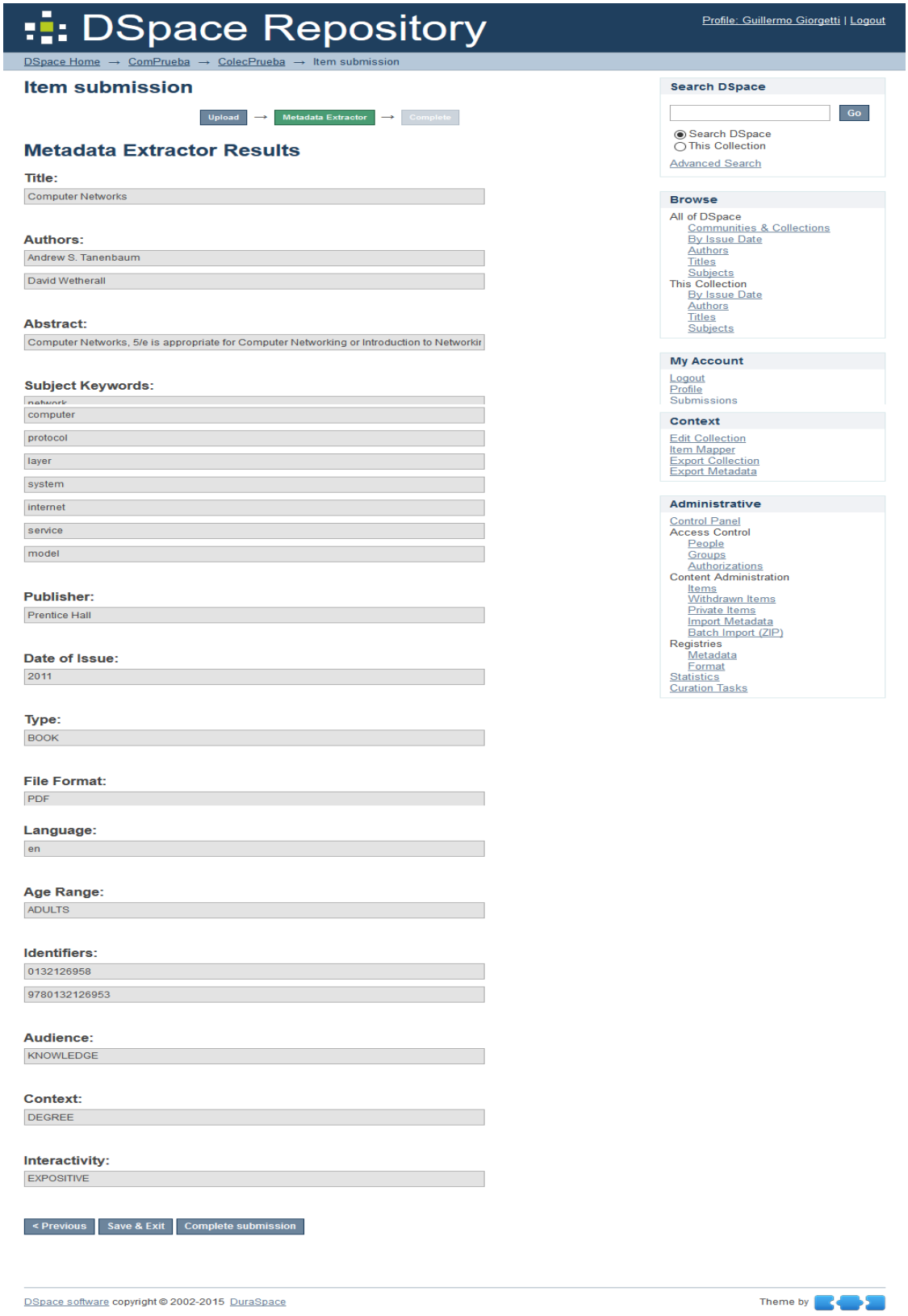

Fig. 4.20. Resultados AMELOIR para libro

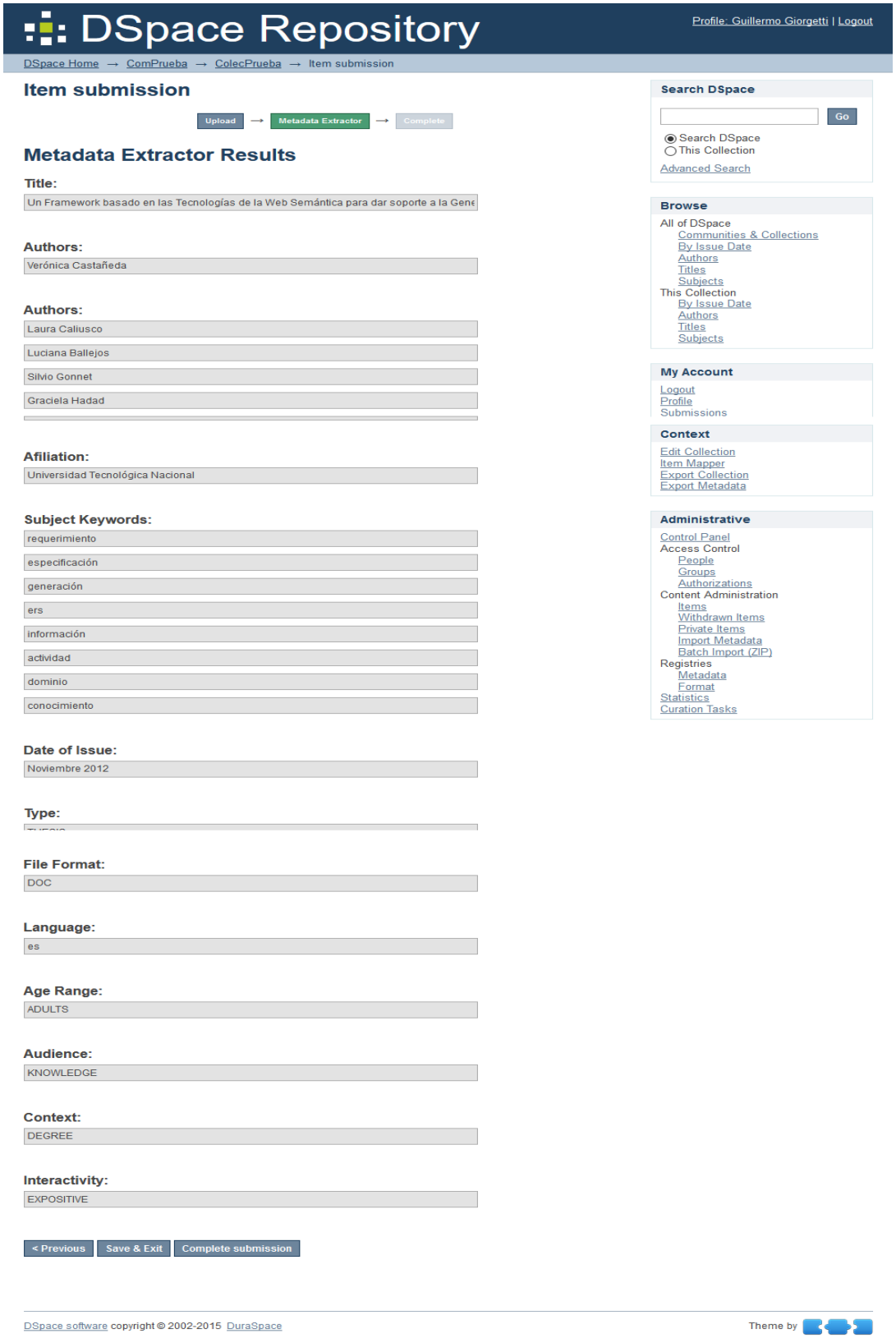

Fig. 4.21. Resultados AMELOIR para tesis de posgrado

Capítulo 5

Conclusiones y Trabajos Futuros

5.1 Conclusiones

En este trabajo se presentó el contexto general en lo que respecta a repositorios institucionales de acceso abierto, considerando los conceptos de objetos de aprendizaje (denominados OA), metadatos y estándares. De igual forma, por tratarse de un trabajo conjunto entre profesionales de Argentina y Colombia, se describió la historia y situación actual de los repositorios institucionales teniendo como punto de referencia aquellos más representativos de ambos países. El análisis efectuado permitió evidenciar que cada vez toma mayor fuerza e importancia esta realidad que se apoya en las tecnologías de la información para innovar en la comunidad educativa, involucrando de manera particular a docentes, estudiantes e investigadores. Con esto, también surge la problemática de la sub-utilización de estas poderosas herramientas, en la mayoría de los casos por desconocimiento en cuanto al uso, la falta de buscadores apropiados y dificultades técnicas que se presentan en la comunidad educativa que implica la atención a distintos niveles de usuarios generalmente no-técnicos; dentro de estas dificultades se encuentra también el entendimiento de los metadatos como descriptores apropiados para encontrar, gestionar, reusar y almacenar OA's en forma efectiva.

Una de las grandes ventajas es el avance que se tiene en cuanto a la definición de estándares para metadatos, que facilitan, de alguna manera, tanto el almacenamiento como la recuperación de OA en los repositorios. Aquí cabe destacar la importancia que tiene el uso de dichos estándares en la definición de nuevos algoritmos para la extracción automática de metadatos, poniendo especial atención a aquellos que se consideran educacionales. Los estándares también permiten lograr una mayor compatibilidad e interoperabilidad entre los diferentes repositorios institucionales que existen, ampliar el rango de búsqueda y la reutilización de los OA, y difundir en mayor medida el acceso abierto al universo de conocimiento científico-educativo que se encuentra disponible en las diferentes instituciones

educativas y en Internet y que pueden llegar a ser muy valiosas y útiles dentro del proceso de enseñanza-aprendizaje.

A partir del análisis comparativo que se realizó de los sistemas extractores de metadatos, se evidenció que en lo que respecta a este tipo de tecnología es mucho el camino que queda por recorrer, no sólo en el momento de autoarchivar y clasificar OA's, sino también en el de ser incluidos en los resultados de las búsquedas, precisamente porque los metadatos asociados a los mismos no son claramente identificados.

Teniendo en cuenta estos aspectos, se diseñó AMELOIR, un nuevo algoritmo para la extracción de metadatos que se pretenda sea incorporado al RI del CIDISI, con el fin último de contribuir de manera significativa a la difusión del conocimiento de esta comunidad educativa. Para esta propuesta se tuvieron en cuenta 6 colecciones predeterminadas a modo de categorías para la clasificación y almacenamiento de los OA's y la extracción de 10 metadatos obligatorios y 6 metadatos descriptivos opcionales compatibles con los estándares Dublin Core e IEEE LOM.

Con AMELOIR, la extracción automática de metadatos se debe ejecutar durante el proceso de carga de un documento al repositorio bajo la plataforma DSpace, sin interacción con el usuario; esto con tal de evitar en lo posible la validación/corrección de la información asociada al OA por parte del usuario que está realizando el proceso de autoarchivo en el repositorio, para disminuir el grado de imprecisión, baja calidad, inconsistencias y discrepancias de los metadatos. Este punto se convierte en uno de los principales aportes de esta tesis en materia de extracción automática de metadatos y por ende, en el ámbito del acceso abierto como medio de distribución y reutilización del conocimiento científico.

5.2 Trabajos Futuros

A partir del algoritmo de extracción automática de metadatos desarrollado surgen nuevas posibilidades que pueden aprovechar al máximo la utilidad de los metadatos y Repositorios Institucionales de acceso abierto:

- *Búsqueda Inteligente*: el mayor potencial del extractor y una de las principales razones por las que surge la necesidad del mismo, es disponer de información confiable que permita llevar a cabo búsquedas de material educativo basadas en el perfil de usuario. Como se dijo anteriormente, estos sistemas de búsqueda son capaces de seleccionar el material que mejor se adapte a las preferencias o necesidades del usuario,

devolviendo así información de mayor precisión y calidad. Esto ayudaría a un usuario a encontrar los recursos educativos que le sean más apropiados de acuerdo con su perfil.

- *Soporte a mayor cantidad de formatos*: actualmente, el extractor tiene en cuenta los formatos de documentos más comunes. Sin embargo, a medida que el Repositorio siga creciendo en contenido almacenado, posiblemente surgirá la necesidad de dar soporte a otros formatos, como Excel, PPT e incluso imágenes y archivos de audio. Teniendo esto en mente, AMELOIR se encuentra desarrollado de manera modularizada, con el objetivo de permitir la fácil extensión del mismo y la reutilización de las funcionalidades disponibles.

- *Aprendizaje Automático*: un punto de investigación para futura mejoras es buscar la forma de obtener resultados más precisos, dotando al sistema de la capacidad de autoaprendizaje. A partir del resultado de extracción de metadatos obtenido y la retroalimentación del usuario, el agente podría establecer métodos de ponderación y evaluación de los metadatos extraídos automáticamente, para así garantizar de cierta manera la calidad de la información que se está extrayendo y almacenando en los repositorios. Esto facilitaría la evolución de los algoritmos de extracción automática de metadatos, y por consiguiente, mejoraría los resultados de las búsquedas que se realizan.

- *Mayor acceso a la información*: sería interesante empezar a explorar el área del reconocimiento de voz y el uso de colores y estructuras adecuadas, apoyados en el uso de inteligencia artificial, como medio para ampliar la accesibilidad a los objetos digitales educativos hacia aquellos segmentos de la población que cuentan con alguna condición especial en lo que respecta a su condición física y su adaptación frente a un computador. Todo ello debe ser tenido en cuenta para no restringir o limitar las bondades del acceso abierto al conocimiento como medio de difusión masivo y universal.

Bibliografía

Adaptive Technology Resource Centre, A. (2002). *ATutor*. Obtenido de http://www.atutor.ca/

Alfano, M., Lenzitti, B., & Visalli, N. (2007). SAXEF : A System for Automatic eXtraction of E-learning object Features. *Je-LKS: Journal of e-Learning and Knowledge Society*, 83-92.

Apache. (2002). *Apache POI - the Java API for Microsoft Documents*. Obtenido de https://poi.apache.org/

Apache. (2011). *Apache Lucene Core*. Obtenido de https://lucene.apache.org/core/

Association for Information Science and Technology, a. (2014). *Hans Peter Luhn*. Obtenido de Hans Peter Luhn: https://www.asist.org/pioneers/hans-peter-luhn/

Astudillo, G. J. (Septiembre de 2011). *Análisis del estado del arte de los objetos de aprendizaje.* La Plata.

Baca, M. (1998). *Introduction to metadata. Pathways to Digital Information.* Los Ángeles, California: Getty Information Institute.

Barkman, P., Brown, D., Brusilovsky, P., Burke, J. R., Fore, M., Hyde, J., . . . Peoples, B. (2002). Draft Standard for Learning Object Metadata. 1-44.

Beel, J., Gipp, B., Langer, S., Genzmehr, M., Wilde, E., Nürnberger, A., & Pitman, J. (2011). Introducing Mr. DLib, a Machine-readable Digital Library. *In Proceedings of the 11th ACM/IEEE Joint Conference on Digital Libraries (JCDL'11)*.

Berry, M. W., & Kog, J. (2010). *Text Mining: Applications and Theory.* John Wiley & Sons.

Billhardt, H. (12 de Diciembre de 2007). Capítulo 7: Expresiones Regulares.

Blanco Suárez, S. (17 de Marzo de 2006). *Santiago Blanco Suárez*. Obtenido de Universidad de Valladolid: https://www.infor.uva.es/~sblanco/Tesis/Metadatos.pdf

Blázquez Ochando, M. (2013). *Técnicas avanzadas de recuperación de información: procesos, técnicas y métodos.* Madrid: Monografías Electrónicas.

Budapest Open Access Initiative. (2002). Obtenido de Budapest Open Access Initiative: http://www.budapestopenaccessinitiative.org/read

Campo Saavedra, M. F., Martínez Barrios, P. D., Ruíz Rodgers, N., & Rendón Osorio, H. J. (2012). *Recursos Educativos Digitales Abiertos COLOMBIA.* Bogotá, D.C: Ministerio de Eduación Nacional.

Casali, A., Deco, C., Bender, C., Fontanarrosa, S., & Sabater, C. (2013). Asistente para el depósito de objetos en repositorios con extracción automática de metadatos. Madrid, España: XV Simposio Internacional de Tecnologías de la Información y las Comunicaciones en la Educación (SINTICE 2013) .

Casali, A., Gerling, V., Deco, C., & Bender, C. (2009). *Un Sistema inteligente para asistir la búsqueda personalizada de objetos de aprendizaje*. Rosario, Santa Fe, Argentina: Universidad Nacional de Rosario.

Casali, A., Gerling, V., Deco, C., & Bender, C. (2011). Recommender System for Personalized Retrieval of Learning Objects. En *Educational Recommender Systems and Technologies: Practices and Challenges* (págs. 182-210).

Chang, C.-C., & Lin, C.-J. (2011). *LIBSVM -- A Library for Support Vector Machines*. Obtenido de https://www.csie.ntu.edu.tw/~cjlin/libsvm/

Consejo Superior de Investigaciones Científicas, C. (2008). *Ranking Web of Repositories*. Obtenido de Ranking Web of Repositories: http://repositories.webometrics.info/es/

Councill, I. G., Giles, C. L., & Kan, M.-Y. (10 de Marzo de 2008). ParsCit: An open-source CRF reference string parsing package.

DCMI. (1995). *DCMI Metadata Basics*. Obtenido de http://dublincore.org/metadata-basics/

DCMI. (1995). *Dublin Core Metadata Element Set, Version 1.1*. Obtenido de http://dublincore.org/documents/dces/index.shtml

DCMI. (1995). *Dublin Core Metadata Initiative*. Obtenido de http://www.dublincore.org/

DCMI-IEEE-MOU. (2000). *Memorandum of Understanding between the Dublin Core Metadata Initiative and the IEEE Learning Technology Standards Committee*. Obtenido de http://dublincore.org/documents/dcmi-ieee-mou/

De Volder, C. (2012). El acceso abierto en Argentina. *Boletín electrónico ABGRA*, 10.

Defense Technical Information Center, D. (2007). *Public Scientific and Technical Information Network*. Obtenido de Public Scientific and Technical Information Network: http://www.dtic.mil/dtic/

Dueñas Fernández, R. A. (Septiembre de 2013). Extracción de información y conocimiento de las opiniones emitidas por usuarios de los sistemas WEB 2.0. Chile: Universidad de Chile.

DuraSpace. (2002). *DSpace*. Obtenido de Dspace: http://www.dspace.org/

Flynn, P., Zhou, L., Maly, K., Zeil, S., & Zubair, M. (2007). Automated template-based metadata extraction architecture. *Asian Digital Libraries*, 327-336.

Frank, E., & Medelyan, O. (1999). *KEA Automatic Keyphrase Extraction*. Obtenido de KEA Automatic Keyphrase Extraction: http://www.nzdl.org/Kea/index_old.html

Giorgetti, C., Romero, L., & Gutierrez, M. (2015). Definición de Metadatos Educativos para Repositorios. *XXI Congreso Argentino de Ciencias de la Computación*, (págs. 398-407). Junín.

Google. (1998). Obtenido de Google: https://www.google.com

Han, H., Lee, C. G., Manavoglu, E., Zha, H., Zhang, Z., & Fox, E. A. (2003). Automatic Document Metadata Extraction using Support Vector Machines. *Proceedings of the 3rd ACM/IEEECS Joint Conference on Digital Libraries*, 37-48.

Hernández, M., & Gómez, J. (2013). Aplicaciones de Procesamiento de Lenguaje Natural. *Revista Politécnica, 32*, 87-96.

Hetzner, E. (2008). A simple method for citation metadata extraction using Hidden. *JCDL '08: Proceedings of the 8th ACM/IEEE-CS joint conference on Digital libraries*, 280-284.

Honorable Cámara de Diputados de la Nación, A. (2010). PROYECTO DE LEY. Creación de Repositorios Digitales Abiertos de Ciencia y Tecnología. CIENCIA ABIERTA ARGENTINA 2010. Argentina.

IBM. (2001). *IBM LanguageWare Resource Workbench*. Obtenido de IBM LanguageWare Resource Workbench: https://www.ibm.com/developerworks/community/groups/service/html/communityview?communityUuid=6adead21-9991-44f6-bdbb-bafod2e8a673&lang=en

IBM. (2009). *AlchemyAPI*. Obtenido de AlchemyAPI: http://www.alchemyapi.com/

IBM. (s.f.). *IBM LanguageWare Resource Workbench*. Obtenido de https://www.ibm.com/developerworks/community/groups/service/html/communityoverview?communityUuid=6adead21-9991-44f6-bdbb-bafod2e8a673

IEEE. (2002). *IEEE LOM*. Obtenido de IEEE LOM: http://www.ieee.org

ISO/IEC. (2003). 11179-3:2003 Information technology. Metadata registries (MDR). *Part 3: Registry metadatamodel and basic attributes*.

Java. (2009). *Apache PDFBox - A Java PDF Library*. Obtenido de Apache PDFBox - A Java PDF Library: http://pdfbox.apache.org/

JBoss, C. (2001). *Drools*. Obtenido de http://www.drools.org/

JBoss, C. (2006). *Drools*. Obtenido de Drools: https://www.drools.org/

JetBrains. (2001). *IntelliJ IDEA*. Obtenido de https://www.jetbrains.com/idea/

Johnson, R. (2002). *Spring Framework*. Obtenido de https://spring.io/

Klink, S., Dengel, A., & Kieninger, T. (2000). Document structure analysis based on layout and textual features. *Proc. of Fourth IAPR International Workshop on Document Analysis Systems*, 99-111.

Learning, A. D. (2000). *SCORM – ADL Net*. Obtenido de SCORM – ADL Net: http://www.adlnet.org/scorm/

Li, Y., Dorai, C., & Farrell, R. (2005). Creating MAGIC: system for generating learning object metadata for instructional content. *Proceedings of the 13th annual ACM international conference on Multimedia*, 367-370.

López Guzmán, C. (2005). Los Repositorios de Objetos de Aprendizaje como soporte a un entorno e-learning. *Trabajo de Grado*. Salamanca, España: Universidad de Salamanca.

Lowagie, B. (2000). *iText Software*. Obtenido de iText Software: http://www.itextpdf.com/

Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S. J., & McClosky, D. (2014). The Stanford CoreNLP Natural Language Processing Toolkit. *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, 55-60.

Marzal García-Quismondo, M. Á., Calzada Prado, J., & Cuevas Cerveró, A. (2006). Desarrollo de un esquema de metadatos para la descripción de recursos educativos: el perfil de aplicación mimeta. *Revista Española de Documentación científica*, 551-571.

Maven. (2002). *Apache Maven Project*. Obtenido de https://maven.apache.org/

Melero, R. (2007). Tendencias, impacto y actitudes entre los investigadores respecto al acceso abierto a las publicaciones científicas (Open Access). *(IATA) Comunicaciones congresos*, 3-7.

Ministerio de Ciencia, T. e. (2002). *Sistema Nacional de Repositorios Digitales - República Argentina*. Obtenido de Sistema Nacional de Repositorios Digitales - República Argentina: http://repositorios.mincyt.gob.ar/

Motz, R., Badell, C., Barrosa, M., Sum, R., Díaz, G., & Castro, M. (2009). LooKIng4LO Sistema Informático para la extracción automática de Objetos de Aprendizaje. En C. Vaz de Carvalho, R. Azambuja Silveira, & M. Caeiro Rodriguez, *TICAI2009:*

TICs para a Aprendizagem da Engenharia (págs. 7-12). IEEE, Sociedade de Educação: Capítulos Espanhol e Português.

Nagy, G., Seth, S., & Viswanathan, M. (Julio de 1992). A prototype document image analysis system for technical journals. IEEE.

NISO. (2012). *JATS: Journal Article Tag Suite.* Obtenido de http://www.niso.org/apps/group_public/download.php/10904/z39.96-2012.pdf

NLM, J. (s.f.). *Archiving and Interchange Tag Set*. Obtenido de http://dtd.nlm.nih.gov/archiving/

Nuance, C. (2000). *OmniPage*. Obtenido de http://www.nuance.com/for-business/by-product/omnipage/index.htm

O'Gorman, L. (1993). The document spectrum for page layout analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1162-1173.

Oracle. (1995). *Java*. Obtenido de https://www.java.com/es/

Padró, L., & Stanilovsky, E. (2012). FreeLing 3.0: Towards Wider Multilinguality. *Language Resources and Evaluation Conference (LREC 2012) ELRA*. Obtenido de http://nlp.lsi.upc.edu/freeling/node/1

PDF, L. (2004). *PDF tk*. Obtenido de https://www.pdflabs.com/tools/pdftk-the-pdf-toolkit/

Pinilla Gómez, A. C., Gutiérrez, M., & Ballejos, L. (2014). EXTRACCIÓN AUTOMÁTICA DE METADATOS A PARTIR DE OBJETOS DE APRENDIZAJE EN UN REPOSITORIO INSTITUCIONAL: ESTADO DEL ARTE. *Anales de las 43 JAIIO*.

Polsani, P. R. (2006). Use and Abuse of Reusable Learning Objects. *Journal of Digital Information, 3*(4).

Porter, M. (1980). An algorithm for suffix stripping. 313-316.

PostgreSQL, G. D. (1996). *PostgreSQL*. Obtenido de https://www.postgresql.org

Referencia, L. (Enero de 2013). América Latina pasa la primera página en Acceso Abierto. *La Referencia visibilizando la ciencia*. América Latina.

Reusable eLearning Object Authoring & Delivery, R. (2004). *Reusable eLearning Object Authoring & Delivery*. Obtenido de http://www.reload.ac.uk/editor.html

Roa, J., Gutiérrez, M., & Stegmayer, G. (2008). FAIA: Framework para la enseñanza de agentes en IA. *Revista Iberoamericana de Informática Educativa*, 43-56.

San Martin, P. S. (2010). PICTO-CIN 0143. *Hacia el desarrollo y utilización de Repositorios de Acceso para Objetos Digitales Educativos en el contexto de las universidades públicas de la región centro-este de Argentina*. Argentina.

San Martin, P., Guarnieri, G., & Bongiovani, P. (2014). PROPUESTA SOCIOTECNOLÓGICA PARA EL DESARROLLO DE REPOSITORIOS DE ACCESO ABIERTO ADECUADOS AL CONTEXTO UNIVERSITARIO ARGENTINO. *e-Ciencias de la Información*. Obtenido de http://dx.doi.org/10.15517/eci.v4i2.15131

Snowtide, I. (2001). *PDFxStream*. Obtenido de https://www.snowtide.com/

Sosa, R., Rodríguez, A., & Motz, R. (2006). Adquiriendo Metadatos para Objetos de Aprendizaje. *First Latin American Conference on Learning Objects (LACLO 2006)*.

Sutton, C., & McCallum, A. (2006). An Introduction to Conditional Random Fields for Relational Learning.

Tanenbaum, A. S., & Wetherall, D. (2011). *Computer Networks* (Quinta ed.). Pearson Prentice Hall. Obtenido de https://books.google.com.ar/books?id=I764bwAACAAJ&printsec=frontcover&dq=editions:ISBN0133499456

Tkaczyk, D., Szostek, P., Dendek, P. J., Fedoryszak, M., & Bolikowski, L. (2014). CERMINE - Automatic extraction of metadata and references from scientific literature. *Proceedings - 11th IAPR International Workshop on Document Analysis Systems, DAS 2014*, 217-221.

Tkaczyk, D., Szostek, P., Dendek, P. J., Fedoryszak, M., & Bolikowski, L. (2014). *GROTOAP2 — The Methodology of Creating a Large Ground Truth Dataset of Scientific Articles*. Obtenido de GROTOAP2 — The Methodology of Creating a Large Ground Truth Dataset of Scientific Articles: http://www.dlib.org/dlib/november14/tkaczyk/11tkaczyk.html

Vilca, O. (2014). Expresiones regulares y autómatas finitos.

Wai Yuen, T. (2007). *AUTOMATIC EXTRACTION OF LEARNING OBJECT METADATA (LOM) FROM HTML WEB PAGE.* HONG KONG.

Walnes, J. (2004). *XStream*. Obtenido de http://x-stream.github.io/

Wiley, D. (2001). Connecting learning objects to instructional design theory: A definition, a metaphor, and a taxonomy. En D. A. Wiley, *The instructional use of learning objects* (págs. 1-35). Logan: Utah State University.

Witten, I. H., Paynter, G. W., Frank, E., Gutwin, C., & Nevill-Manning, C. G. (1999). KEA: Practical Automatic Keyphrase Extraction.

Zyl, J. v. (2002). *MVNRepository*. Obtenido de https://mvnrepository.com/

Anexos

Anexo 1. Repositorios Institucionales representativos de Argentina

SEDICI - Repositorio Institucional http://sedici.unlp.edu.ar/			
Institución: Universidad Nacional de La Plata (UNLP) **Plataforma:**DSpace **Estándar**: DCMI **Ranking a Nivel Nacional:** 1 **Ranking a Nivel Latinoamérica:** 5 **Ranking a Nivel Mundial:** 78			
OPCIONES DE BÚSQUEDA	**CATEGORÍAS/COLECCIONES**	**COINCIDENCIAS Y CONSISTENCIAS EN LAS BÚSQUEDAS**	**OBERVACIONES**
Tres grandes agrupaciones: **colecciones**, **autores** y **temas**. Dentro de **colecciones** se encuentra una sub-clasificación así: - Academia Nacional de Agronomía y Veterinaria (ANAV). - Biblioteca Digital. - Eventos. - Red de universidades con carreras en informática (RedUNCI). - Revistas. - Unidades académicas. Así mismo, el contenido de cada colección lo clasifican por: tipo de documento, fecha de publicación, materia, autor, palabra clave.	**Tesis**: tesis de grado, post-grado y otros documentos. **Revistas**: publicaciones en revistas científicas. **Eventos**: ponencias realizadas en congresos y conferencias. **Libros**: libros digitalizados y e-books. **Red UNCI**: artículos y ponencias de la red UNCI. **Radio Universidad**: entrevistas y producción artísticas, entre otros audios. **Emergencia hídrica**: trabajos dentro del Plan de Gestión Integrada de Riesgos de Desastres. Todos los tipos de documentos que se pueden encontrar en SEDICI se encuentran especificados en http://sedici.unlp.edu.ar/pages/FAQ#tipos	Todo lo referente a búsquedas se encuentra en http://sedici.unlp.edu.ar/pages/FAQ#busqueda: Para realizar una búsqueda simple basta introducir una *palabra* (por ejemplo, "física") en el cuadro de búsqueda de la página principal de SEDICI. El sistema recuperará *todos los registros que contengan la palabra "física" en cualquiera de sus metadatos* (aquellos datos que sirven para catalogar correctamente un recurso). Sin embargo, este tipo de búsqueda puede traer resultados no deseados o no del todo atinados. Para refinarla, puede realizarse entonces una búsqueda avanzada, en la que se especifiquen otros criterios además de la palabra deseada. Estos criterios pueden incluir: *si la palabra deseada se encuentra en el título o si se prefiere buscar por fecha de publicación*. También puede utilizarse la barra de navegación de la derecha, donde aparecen reflejados los resultados de la búsqueda de acuerdo a los siguientes criterios: *tipo de documento* (cuáles y qué cantidad contienen la palabra buscada); *fecha de publicación* (rango de años y qué cantidad contienen la palabra buscada); *materia*; *autor* (qué autores y qué cantidad de trabajos por autor tienen la palabra buscada) y *palabra-clave*. Otro tipo de búsquedas también son posibles en SEDICI: por ejemplo, si se hace clic en el *nombre de un autor*, el sistema recuperará *todos los recursos donde ese autor figure (no necesariamente en calidad de autor, ya que también puede hacerlo en calidad de director de tesis, editor, etc.)*; del mismo modo, al hacer clic en un *descriptor o una palabra-clave*, el sistema recuperará *todos los recursos que los incluyan*. Asimismo, es posible buscar en todo SEDICI o sólo en una *determinada comunidad o colección*.	De acuerdo a la política de metadatos (http://sedici.unlp.edu.ar/pages/politicas) "Todos los registros de metadatos de las obras depositadas en el repositorio son diseminadas a partir de protocolos de interoperabilidad bajo formato DublinCore y similares".

Biblioteca digital UNCuyo http://bdigital.uncu.edu.ar/		
Institución: Universidad Nacional de Cuyo (UNCUYO) **Ranking a Nivel Nacional:** 3 **Ranking a Nivel Latinoamérica:** 29 **Ranking a Nivel Mundial:** 558		
OPCIONES DE BÚSQUEDA	**CATEGORÍAS/COLECCIONES**	**COINCIDENCIAS Y CONSISTENCIAS EN LAS BÚSQUEDAS**
Búsqueda general: expresión a buscar (palabra clave) *Búsqueda por tipo*: compuesta por: - Texto - Video - Audio *Disciplinas:* materias Los documentos contenidos en esta Biblioteca Digital pueden consultarse de diversas maneras (http://bdigital.uncu.edu.ar/paginas/guia-de-uso.php): Búsqueda por palabras clave generales, lo que arrojará una variedad de resultados sin discriminación de formatos (ofrecerá resultados de texto, audio y video) Búsqueda avanzada: Permite buscar sólo documentos de un determinado formato, y trabajar con operadores de búsqueda. Búsqueda por tipos de documento: El sistema ofrece navegar los diversos formatos (texto, audio y video) a través de sus distintos tipos de documento. Esto es en Textos > Revistas, Tesis, Informes de Investigación y Libros; en Videos y Audios > Documentales, Informes especiales, Eventos académicos y Entrevistas. Por disciplina: Para una navegación temática, el sistema ofrece una navegación a través de las disciplinas de estudio, según una categorización de la SECyT.	Se denominan Tipo: - **Texto**: revistas, informes de investigación, libros electrónicos, producción científica académica, producción científica independiente, tesis, memoria académica, material didáctico. - **Video**: archivos audiovisuales (entrevistas, formación académica, documentales, noticias universitarias), microprogramas. - **Audio**: archivos sonoros (documentales), microprogramas sonoros, entrevistas, eventos académicos, informes especiales.	*Estructura de búsqueda* agua vino se buscan las dos palabras usando un operador "Y" también puede usarse agua +vino obteniendo el mismo resultado. agua -vino : se busca la primer palabra y se excluyen las que tengan la segunda usando el operador "-" *Uso de singulares y plurales* efluentes es buscado igual que efluente ya que se singularizan las palabras para realizar las búsquedas. *Uso de acentos* se puede evitar el uso de acentos al buscar contaminacion y contaminación arrojan los mismos resultados. *Otras opciones* Siempre los resultados pueden filtrarse por el formato del objeto: Texto, video o Audio
Biblioteca Virtual UNL http://bibliotecavirtual.unl.edu.ar/		

Institución: Universidad Nacional del Litoral (UNL)
Plataforma:DSpace
Estándar: DCMI
Ranking a Nivel Nacional: 2
Ranking a Nivel Latinoamérica 19
Ranking a Nivel Mundial: 486

OPCIONES DE BÚSQUEDA	CATEGORÍAS/COLECCIONES	OBERVACIONES
Se divide por bibliotecas, así: - *Tesis:* es un repositorio que cuenta con las tesis en formato digital producidas en las Unidades Académicas (o facultades) de la Universidad Nacional del Litoral (UNL). - *Publicaciones Periódicas*: contiene las colecciones digitalizadas de las publicaciones periódicas publicadas por la Unidades Académicas (o facultades) y otros organismos de la Universidad Nacional del Litoral (UNL). Estas publicaciones son editadas por el Centro de Publicaciones de la Universidad Nacional del Litoral, el cual es el organismo encargado de las ediciones en la Universidad del material impreso y digital de obras científicas, técnicas y artísticas. - *Material Educativo*: es un repositorio que cuenta con recursos digitales para ser utilizados como herramientas en el aprendizaje de las diferentes asignaturas de las carreras de la Universidad Nacional del Litoral. - *Imágenes*: repositorio permanente de las fotografías obtenidas de las actividades académicas y de gestión de la Universidad Nacional del Litoral (UNL). - *Colecciones:* contiene Fondos que reúnen las obras más representativas de docentes e investigadores destacados de la Universidad en todas las ramas de la ciencia, donaciones por parte de personas e instituciones con proyección pública y otro material bibliográfico antiguo en temas diversos existente en la Bibliotecas de la Universidad, que por razones de conservación no puede ser accedido por el público en su soporte original. A su vez, *Tesis* y *Publicaciones Periódicas* se subdividen en unidades académicas (facultades), dentro de las cuales aparece un filtro inicial por los diferentes posgrados asociados a cada una de ellas, o las opciones de buscar por Palabra Clave o listar por Títulos, Autores, Temas o por Fecha. La búsqueda avanzada está compuesta por la búsqueda de palabras a nivel de Autor, Título, Palabra Clave o Resumen, contando con la posibilidad de hacer combinaciones de, a lo máximo, 3 criterios. En cuanto a *Material Educativo*, está el filtro inicial de buscar por las materias de Física o Química, y dentro de cada una la opción de búsqueda por Palabra Clave o listar por Autor, Título o Materia. La búsqueda avanzada está compuesta por la búsqueda de palabras a nivel de Palabra Clave, Autor, Título, Tema, Resumen, Colección, Sponsor, Identificador y Lengua, contando con la posibilidad de hacer combinaciones de, a lo máximo, 3 criterios. En la Biblioteca de *Imágenes* (Fototeca), existe un filtro inicial por: - Edificios y Lugares - Eventos y Temas - Personas Al ingresar a una de éstas, se muestra la opción de búsqueda por Palabra Clave o listar por Autor, Título o Palabra Clave; también se muestra un listado de las colecciones que la componen, así: - Edificios y Lugares: aulas, ciudad universitaria, facultades y escuelas, laboratorios, planta de alimentos, predio UNL-ATE, rectorado. - Eventos y Temas: académicos, científicos, culturales, institucionales. - Personas: decanos, rector. En esta biblioteca no existe la opción de búsqueda avanzada. Por último, en la biblioteca de *Colecciones* existe un único filtro inicial por el Fondo Hugo Gola, dentro del cual aparece un filtro inicial por las colecciones El Poeta y su Trabajo y Poesía y Poética, o las opciones de buscar por Palabra Clave o listar por Títulos, Autores, Temas. La búsqueda avanzada está compuesta por la búsqueda de palabras a nivel de Autor, Título, Palabra Clave o Resumen, contando con la posibilidad de hacer combinaciones de, a lo máximo, 3 criterios.	Corresponden a las bibliotecas, es decir: - Tesis - Publicaciones Periódicas - Material Educativo - Imágenes - Colecciones	No hace uso del sistema de autoarchivo, sino que para publicar en el repositorio, el recurso debe ser enviado previamente al personal de la UNL para su revisión y clasificación.

Rephip – Repositorio Hipermedial http://rephip.unr.edu.ar/		
Institución: Universidad Nacional de Rosario (UNR) **Plataforma:**DSpace **Estándar**: DCMI **Ranking a Nivel Nacional:** 6 **Ranking a Nivel Latinoamérica:** 60 **Ranking a Nivel Mundial:** 817		
OPCIONES DE BÚSQUEDA	**CATEGORÍAS/ COLECCIONES**	**COINCIDENCIAS Y CONSISTENCIAS EN LAS BÚSQUEDAS**
Existe un primer filtro de búsqueda por: - Comunidades y colecciones. - Fecha de publicación - Autores - Títulos - Temas Dentro de la búsqueda avanzada se contemplan los siguientes "tipos de búsqueda": - Ámbito de búsqueda (comunidad) - Texto completo - Resumen - Series - Autor - Título - Palabra clave - Idioma (ISO) - Tipo MIME (tipo de archivo) - Sponsor - Identificación	Las colecciones se presentan de acuerdo a la comunidad.	Cuando se efectúa una "Búsqueda a texto completo", es posible que los resultados se muestren incluso clasificados por colecciones e ítems (como sucede cuando se busca "sistemas inteligentes"); este tipo de búsqueda se realiza sobre el texto completo, tal como su nombre lo indica, incluyendo resumen. La palabra clave también se busca a lo largo del texto completo. La búsqueda por Idioma se realiza conforme a ISO, sin embargo, al colocar "ita", me trae un texto en español (al parecer por mala clasificación). La búsqueda por título se realiza buscando todos los títulos que contengan la palabra o palabras ingresadas. En tal caso, no es posible hacer un búsqueda por título completo y que me traiga un único resultado. Ejemplo: al ingresar la búsqueda por título "vivencias del camino de Santiago", el primer resultado efectivamente es el objeto cuyo título coincide exactamente con el parámetro de búsqueda, pero me muestra otros 803 resultados, cuyos títulos contienen alguna de las palabras del criterio de búsqueda. La búsqueda por tipo MIME no funciona adecuadamente, quizás debido a la mala calidad de los metadatos de los OA. Por ejemplo, al buscar por "PDF" trae 477 resultados, que seguramente son los únicos que tienen este metadato completo y bien diligenciado. Al realizar la búsqueda por Sponsor, por ejemplo "UNR" trae algunos resultados. La búsqueda por Autor se puede hacer por nombre y/o apellido. La búsqueda por Identificador corresponde al ISBN.

Corciencia - Repositorio Digital de Investigaciones Científicas y Tecnológicas de Córdoba http://www.corciencia.org.ar/			
Institución: Gobierno de la Provincia de Córdoba **Plataforma:** EPrints3 **Estándar**: DCMI			
OPCIONES DE BÚSQUEDA	**CATEGORÍAS/COLECCIONES**	**COINCIDENCIAS Y CONSISTENCIAS EN LAS BÚSQUEDAS**	**OBERVACIONES**
Se puede navegar por: - Año - Descriptor (tema) - División (Comunidad) - Autor La opción de búsqueda avanzada incluye: - Documento - Título - Autores - Abstract - Fecha - Términos no controlados (palabras clave) - Descriptores - Tipo de ítem - Departamento - Editores - Estado: enviado, publicado - Referenciado - Diario o título de la publicación	Se denominan Tipo de Ítem: - Proyectos de investigación - Artículo - Sección de libro - Reseña bibliográfica - Libro - Patente o marca - Imagen - Video - Audio - Otro	En la navegación por año, existe una categoría UNESPECIFIED, que corresponde a aquellos OA que no tiene el metadato de año. Descriptor = Tema División = Comunidad Términos no controlados = Palabras clave Algunos criterios de búsqueda como Departamento no fue posible determinar con exactitud a qué corresponden.	De acuerdo al instructivo de uso, este repositorio no cuenta con un extractor automático de metadatos, ya que todos los datos del OA deben ser ingresados por el usuario que está realizando el proceso de carga. Los metadatos obligatorios son: título, autores, referenciado, estado (enviado, no publicado), diario o título de la publicación, descriptores (permiten delinear el área temática específica de los contenidos, en la opción "Términos no controlados" pueden colocarse palabras importantes que completen la descripción del documento a modo de palabras claves). Tiene la opción de exportar los resultados de las búsquedas en formato DublinCore.

Biblioteca electrónica de ciencia y tecnología http://www.biblioteca.mincyt.gov.ar/index.php			
Institución: Ministerio de Ciencia y Tecnología (MinCyT) **Plataforma:**EBSCO DiscoveryService			
OPCIONES DE BÚSQUEDA	**CATEGORÍAS/COLECCIONES**	**COINCIDENCIAS Y CONSISTENCIAS EN LAS BÚSQUEDAS**	**OBERVACIONES**
Se da la opción de acceso alfabético a libros, acceso por BD y acceso a revistas. Existe tres opciones iniciales: palabra clave, título y autor. Una vez se ejecuta la búsqueda inicial (búsqueda básica), aparecen otros elementos para "depurar la búsqueda" estos son: *Modos y ampliadores de búsqueda*: - Modos de búsqueda: palabra clave/frase, buscar todos mis términos, buscar uno de mis términos, búsqueda inteligente sugerencia. - Otras opciones: aplicar palabras relacionadas, también buscar en el texto completo de los artículos. *Límite de resultados* - Texto completo. - Publicaciones arbitradas. - Nombre de la publicación. - Autor. - Título. - Fecha de publicación desde - hasta: mes/año. - Tipos de vista rápida de imágenes: black and whitephotograph, diagram, illustration, chart, color photograph, graph, map. - Idioma. - Tipo de recurso: todos los resultados, publicaciones académicas, libros, revistas, informes, críticas. - Materia. - Editor. - Publicación. - Lengua. - Geografía. - Proveedor del contenido. También es posible buscar material mediante la combinación de los siguientes elementos (búsqueda avanzada): - TX Alltext. - AU Author. - TI Title. - SU SubjectTerms. - SO JournalTitle/Source. - AB Abstract. - IS ISSN. - IB ISBN.	Se denominan Tipo de Recurso: - Publicaciones académicas - Libros - Revistas - Informes - Críticas	La búsqueda por Título funciona buscando que una palabra esté incluida dentro del título. La búsqueda por Autor se puede hacer por nombre, por apellido o por todo completo. En la búsqueda por Palabras Clave se recomienda utilizar los operadores booleanos AND, OR y NOT. También muestra los resultados obtenidos mediante búsqueda federada (aquellas BD que no están pre-indexadas en el buscador)	Cuenta con un tutorial de búsqueda básica.

BDU² Repositorios Institucionales http://bdu.siu.edu.ar/cgi-bin/reporpt.pl		
Institución: Consorcio de Universidades SIU **Plataforma:** Protocolo Open Archives Initiative		
OPCIONES DE BÚSQUEDA	**CATEGORÍAS/COLECCIONES**	**COINCIDENCIAS Y CONSISTENCIAS EN LAS BÚSQUEDAS**
Búsqueda simple: expresión a buscar (palabra clave) y criterio (todos, autores, título, materias). Los resultados que la búsqueda arroje serán categorizados por diferentes criterios. Si se desea buscar un término truncado puede hacerlo agregando a su expresión el símbolo $ Por ejemplo: democra$ Con lo que aparecerán los términos: Democracia Democráticos Democráticas, etc. *Búsqueda avanzada*: compuesta por: - Combinación de una a 3 expresiones (palabras clave) bajo los criterios todos, autores, título, materias, notas, editor. Se pueden combinar con los diferentes operadores lógicos AND, OR y NOT. - Año de publicación desde-hasta. - Tipo de recurso: más de 30 tipos.	Se denominan Tipo de Recurso, son más de 30.	La búsqueda por Título funciona buscando que una palabra esté incluida dentro del título. La búsqueda por Autor se puede hacer por nombre, por apellido o por todo completo. En la búsqueda por Palabras Clave se recomienda utilizar los operadores booleanos AND, OR, NOT. La búsqueda por Notas funciona de manera similar a la búsqueda por Palabras Clave (Todos) ya que se realiza sobre título, objetivos, materias.

Anexo 2. Repositorios Institucionales representativos de Colombia

Repositorio Institucional UN http://www.bdigital.unal.edu.co/			
Institución: Universidad Nacional de Colombia (UN) **Plataforma:**DSpace **Estándar**: DCMI **Ranking a Nivel Nacional:** 1 **Ranking a Nivel Latinoamérica:** 6 **Ranking a Nivel Mundial:** 96			
OPCIONES DE BÚSQUEDA	**CATEGORÍAS/COLECCIONES**	**COINCIDENCIAS Y CONSISTENCIAS EN LAS BÚSQUEDAS**	**OBERVACIONES**
Consulta general por: - Año - Materia - División: sede, facultad, dirección, etc. - Autor - Tipo de documento *Búsqueda simple* por palabra clave. *Búsqueda avanzada* que incluye: - Documentos - Título - Autores - Resumen - Fecha - Palabras clave - Temática: grandes clasificaciones: 0. Generalidades 1. Filosofía y psicología 2. Religión 3. Ciencias sociales 4. Lenguas 5. Ciencias naturales y matemáticas 6. Tecnología (ciencias aplicadas) 7. Las artes; Bellas artes y artes decorativas 8. Literatura y retórica 9. Geografía e Historia - Unidad administrativa (a nivel de la universidad) - Tipo de documento	Se denominan Tipo de Documento: - Artículo - Capítulo de libro - Documento de trabajo - Ponencia - Libro - Tesis/trabajos de grado - Exhibición - Composición musical - Imagen - Objeto de aprendizaje - Otro	En la búsqueda por Autor al colocar nombre y apellido trae todos los OA cuyo(s) autor(es) tenga el nombre O el apellido introducido en la búsqueda. No trabaja sobre Colaborador/Asesor. La búsqueda por Documentos es similar a la búsqueda por Palabra Clave, pues realiza la búsqueda sobre todo el documento. La búsqueda por Título puede ser por título completo o por palabra clave dentro del título. La búsqueda por Resumen se realiza únicamente sobre aquella sección "Resumen" de los OA. La búsqueda por Fecha se refiere a la fecha de publicación del documento (no de archivo en el repositorio). La búsqueda por Programa trabaja de manera similar a la búsqueda por Temática (ésta última está predeterminada con un listado). Intenté realizar la búsqueda por Editores con diversos criterios pero no me trajo ningún resultado.	Tiene la opción de exportar los resultados de las búsquedas en formato DublinCore.

- Tipo de tesis (maestría, doctorado, pregrado, otra) - Programa académico - Editores - Estado (publicado, en prensa, enviado, no publicado) - Aprobado mediante evaluación por pares - Título de la revista o publicación - Formato			

Repositorio InstitucionalEdocUR

http://repository.urosario.edu.co/

Institución: Universidad del Rosario (UR)
Plataforma:DSpace
Ranking a Nivel Nacional: 2
Ranking a Nivel Latinoamérica: 15
Ranking a Nivel Mundial: 359

OPCIONES DE BÚSQUEDA	CATEGORÍAS/COLECCIONES	COINCIDENCIAS Y CONSISTENCIAS EN LAS BÚSQUEDAS	OBERVACIONES
Búsqueda simple por palabra clave. *Búsqueda avanzada* que incluye: - Título - Autor - Tipo de documento - Tema - Palabra clave - Resumen - Tipo acceso - Lengua También existen listados generales por: - Comunidades: acervo institucional, investigación, producción editorial, tesis y disertaciones académicas, trabajos de grado. - Título - Autor - Tipo de documento - Materia	Se denomina Tipo de Documento: - Artículo - Diario - Documento de conferencia - Documento de trabajo - Documento histórico - Documento institucional - Imagen - Libro - Parte de libro - Reporte - Revista - Tesis de Maestría - Tesis Doctoral - Trabajo de grado - Video	En la búsqueda por Autor al colocar nombre y apellido trae todos los OA cuyo(s) autor(es) tenga el nombre O el apellido introducido en la búsqueda. ((author:Paolaauthor:Bongiovani)) La búsqueda por Título puede ser por título completo o por palabra clave dentro del título. ((title:Argentina)) En la búsqueda por Tipo de Documento se debe introducir alguna de las colecciones que se manejan, ejemplo, "revista". ((type:revista)) La búsqueda por Tema funciona como búsqueda por Palabra Clave. ((keyword:biología)) La búsqueda por Resumen se realiza sobre el "abstract" del documento. ((abstract:sistema)) La búsqueda por Tipo de Acceso no trajo resultados, a pesar de introducir los mismos criterios que se traen en la vista previa de los documentos, por ejemplo, Restringido (Permitido a grupos específicos), Abierto (texto completo) La búsqueda por Lengua se realiza de acuerdo a ISO, ejemplo, SPA, ENG. ((language:SPA))	De acuerdo al instructivo de uso, este repositorio no cuenta con un extractor automático de metadatos, ya que todos los metadatos del OA deben ser ingresados por el usuario que está realizando el proceso de carga. Los metadatos obligatorios dependen del tipo de colección que se está archivando, así: - Trabajo de grado y tesis y disertaciones. - Documentos institucionales. - Investigación. - Producción editorial. También se observa que algunos de los metadatos serán completados por personal de la biblioteca

Repositorio Institucional PUJ http://repository.javeriana.edu.co/			
Institución: Pontificia Universidad Javeriana (PUJ) **Plataforma:**DSpace **Estándar:** DCMI **Ranking a Nivel Nacional:** 3 **Ranking a Nivel Latinoamérica:** 26 **Ranking a Nivel Mundial:** 521			
OPCIONES DE BÚSQUEDA	**CATEGORÍAS/COLECCIONES**	**COINCIDENCIAS Y CONSISTENCIAS EN LAS BÚSQUEDAS**	**OBERVACIONES**
Búsqueda simple por palabra clave. *Búsqueda avanzada* que incluye: - Palabra clave - Autor - Título - Tema - Resumen - Colección - Sponsor - Identificador - Lengua La búsqueda puede ser combinada mediante el uso de operadores booleanos AND, OR, NOT. También existen listados generales por: - Comunidades, subcomunidades y colecciones - Fecha de publicación - Autor - Título - Materia - Serie - Tipo de documento	Se denomina Tipo de Documento: - Sonido - Tesis Doctoral - Trabajo de Grado - Trabajo de Grado Maestría	En la búsqueda por Autor al colocar nombre y apellido trae todos los OA cuyo(s) autor(es) tenga el nombre O el apellido introducido en la búsqueda. ((author:CARLOSauthor:PÉREZ)) La búsqueda por Título puede ser por título completo o por palabra clave dentro del título. ((title:SISTEMA)) La búsqueda por Tema funciona como búsqueda por Palabra Clave. ((keyword:biología)) La búsqueda por Resumen se realiza sobre el "abstract" del documento. ((abstract:inteligente)) La búsqueda por Tipo de Acceso no trajo resultados, a pesar de introducir los mismos colecciones que se manejan en el repositorio; en cambio, por ejemplo, si en la búsqueda por Palabra Clave se introduce como criterio la palabra "tesis", se muestra un listado de "Resultados por colección", con enlaces a cada uno. ((tesis)) No se logró ningún resultado en la búsqueda por Sponsor. La búsqueda por Identificador se realiza sobre el código asignado al OA dentro del repositorio; este aparece en la parte superior del resumen del OA, marcado así: "Por favor, use este identificador para citar o enlazar este ítem: http://hdl.handle.net/10554/1445"; introduciendo como criterio de búsqueda por Identificador "10554/1445", se obtiene el documento con dicho identificador asociado. ((identifier:10554/1445)) La búsqueda por Lengua no trajo resultados, de hecho en el resumen de los datos de los OA no se muestra.	La organización de las Colecciones responde a la estructura definida por BDCOL - Biblioteca Digital Colombiana, con sus correspondientes tipos documentales. De acuerdo al instructivo de uso, este repositorio no cuenta con un extractor automático de metadatos, ya que todos los metadatos del OA deben ser ingresados por el usuario que está realizando el proceso de carga. Los datos requeridos son: título, autor(es), email, directores (si aplica), fecha de publicación y resumen; en otro paso se solicitan las palabras claves del documento en español y en inglés (si aplica), patrocinadores (si aplica), el formato del archivo y seleccionar una de las opciones de confidencialidad. Por último, la información de Materias será ingresada exclusivamente por personal de la biblioteca.

BDCOL – Biblioteca Digital Colombiana http://190.242.114.6/bdcol.html			
Institución: Ministerio de Educación Nacional **Plataforma:**DSpace			
OPCIONES DE BÚSQUEDA	**CATEGORÍAS/COLECCIONES**	**COINCIDENCIAS Y CONSISTENCIAS EN LAS BÚSQUEDAS**	**OBSERVACIONES**
Búsqueda simple por palabra clave. Cuando se ejecuta una búsqueda simple se muestran otros filtros sobre el resultado de la búsqueda que son: institución, tipo, lenguaje y año. *Búsqueda avanzada* que incluye (trae un listado de valores por defecto): - Título - Autor - Tema - Año - Institución - Repositorio	Se denomina Tipo: - General - Objeto de conferencia - Artículo - Libro - Tesis de maestría - Presentación	La búsqueda por Palabra Clave se realiza sobre título, descripción, tema. No se pueden revisar otras coincidencias y consistencias, dado que la búsqueda avanzada trabaja sobre listados de valores por defecto.	Para cada uno de los recursos incluidos en el resultado de una búsqueda se muestra el siguiente detalle: - Título - Autor - Descripción - Tema - Colección - Tipo - Contexto Así mismo, se muestra el link al recurso y el año de publicación, y en algunos casos se muestra un resumen.
Colombia Aprende http://aprende.colombiaaprende.edu.co/es/contenidoslo			
Institución: Ministerio de Educación Nacional **Plataforma:**DSpace **Estándar:** DCMI/SCORM			
OPCIONES DE BÚSQUEDA	**CATEGORÍAS/COLECCIONES**	**COINCIDENCIAS Y CONSISTENCIAS EN LAS BÚSQUEDAS**	
Búsqueda simple por colecciones y recursos (palabra clave). *Búsqueda avanzada* agrupada en primera instancia por Colecciones o Recursos, incluye: - Identificador de la colección - Nombre de la colección - Autor/Colaborador - Palabra clave de la colección - Entidad	Tiene 307 colecciones disponibles que agrupan 27881 recursos.	Para realizar la búsqueda por Entidad se debe ingresar el criterio completo, ejemplo, "Ministerio de Educación Nacional". La búsqueda por Nombre de la Colección puede ser por nombre completo o por palabra clave dentro del nombre. La búsqueda por Palabra Clave se realiza exclusivamente sobre las palabras clave del recurso. La búsqueda por Identificador de la colección se realiza sobre el código asignado al OA dentro	

- Nivel educativo (primera infancia, educación preescolar. básica y media, educación superior, educación para el trabajo y el desarrollo humano) Estos criterios de búsqueda se amplían con los siguientes: - Público al que va dirigido - Estrategia (internet, radio educativa, televisión educativa) - Formato (audio, imagen, vídeo, documento, otro, web) - Tipo de contenido - Dewey - Grupo poblacional		del repositorio; este aparece en la parte superior del resumen del OA, por ejemplo, "Colección 0273". Evidentemente la búsqueda avanzada está más orientada a búsqueda de colecciones que a búsqueda de recursos.

Anexo 3. Resumen de Metadatos definidos por cada Categoría

CATEGORÍA/ METADADO	Term	Libros	Revistas	Producción en extensión	Producción en Tecnología	Informe Investigación	Tesis	Jornadas, Congresos, Conferencias	Producción en Arte	Material Didáctico	Producción Curricular	Fuentes Primarias	Artículos	Capítulo de libro
Fuente de Financiamiento	contributor/lifecycle.contribute					X								
Director	contributor/lifecycle.contribute						X							
Organizador	contributor/lifecycle.contribute							X						
Colaborador/Sponsor	contributor/lifecycle.contribute	X	X	X	X				X	X	X	X	X	X
Compilador	creator/life cycle.contribute.role.author		X											
Editor	creator/life cycle.contribute.role.author								X					
Autor/(es)	creator/life cycle.contribute.role.author	X		X	X	X	X		X	X	X	X	X	X
Fecha de publicación	date/life cycle.contribute.date	X	X	X	X	X	X	X	X	X	X	X	X	X
Resumen	description/general.description	X	X	X	X	X	X	X	X	X	X	X	X	X
Filiación	description/general.description	X	X	X	X	X	X	X	X	X	X	X	X	X
Contexto	educational.context	X	X	X	X	X	X	X	X	X	X	X	X	X
Dificultad	educational.difficulty	X	X	X	X	X	X	X	X	X	X	X	X	X
Tipo de interactividad	educational.interactivity	X	X	X	X	X	X	X	X	X	X	X	X	X
Rango de edad	educational.typicalagerange	X	X	X	X	X	X	X	X	X	X	X	X	X
Formato	format/technical.format	X	X	X	X	X	X	X	X	X	X	X	X	X
ISSN	identifier/general.identifier.		X					X						

	entity													
ID	identifier/general.identifier.entity			X	X	X								
ISBN	identifier/general.identifier.entity	X					X		X	X	X	X	X	X
Idioma	language/general.language	X	X	X	X	X	X	X	X	X	X	X	X	X
Editor	publisher/lifecycle.contribute.role. Publisher	X	X	X	X	X	X	X		X	X	X	X	X
Relación	relation/relation	X	X	X	X	X	X	X	X	X	X	X	X	X
Contenido de los Derechos (texto)	rights	X	X	X		X	X	X	X		X	X	X	
Derecho de acceso	rights/right.description	X	X	X			X	X						
Titular de los derechos	rights/right.description								X	X	X	X	X	X
Fuente	source/relation.resource	X	X	X	X	X	X	X	X	X	X	X	X	X
Palabras Clave	subject/general.keyword	X	X	X	X	X	X	X	X	X	X	X	X	X
Título	title/general.title	X	X	X	X	X	X	X	X	X	X	X	X	X
Tipo	type/educational.learningResourceType	X	X	X	X	X	X	X	X	X	X	X	X	X