EXTRACCIÓN AUTOMÁTICA DE METADATOS A PARTIR DE OBJETOS DE APRENDIZAJE EN UN REPOSITORIO INSTITUCIONAL: ESTADO DEL ARTE

Adriana Pinilla[1], Milagros Gutiérrez[2] y Luciana Ballejos[2],

[1]Facultad Regional Santa Fe - Universidad Tecnológica Nacional
Santa Fe, Argentina
[2]Centro de Investigación CIDISI - Facultad Regional Santa Fe - Universidad Tecnológica Nacional
Santa Fe, Argentina

Resumen

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, 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*.

En general, este trabajo ofrece una perspectiva sobre el estado actual de las investigaciones acerca de extracción automática de metadatos, estableciendo las bases para futuras investigaciones en el marco concreto de objetos de aprendizaje en repositorios institucionales de acceso abierto. Primero se presenta lo relacionado con los estándares para metadatos; luego se elabora un diagnóstico sobre la historia y situación actual de los repositorios institucionales en Argentina y Colombia; posteriormente se evalúan diferentes propuestas actuales de extracción automática de metadatos, a la luz del cumplimiento de estándares y algunos aspectos relevantes para el análisis y diseño de sistemas de extracción automática; finalmente, se presentan las conclusiones y trabajos futuros en el área.

Palabras clave: objetos de aprendizaje, metadatos, estándares, repositorios institucionales, extracción automática, acceso abierto.

Introducción

Los objetos de aprendizaje (de ahora en adelante denominados OA), son "cualquier recurso digital que puede ser reutilizado para la enseñanza" [Wiley, 2002]; estos objetos pueden adquirir formas muy diversas y presentarse en diferentes formatos y

soportes. Además, deben tener ciertas características, entre las cuales, las más significativas son: *accesibilidad*, *reusabilidad* e *interoperabilidad* [Polsani, 2003].

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. En general, los metadatos son un conjunto de atributos o etiquetas que describen las principales características de un OA y proporcionan información adicional sobre el mismo, fundamental para garantizar el éxito en la interconexión entre repositorios y facilitar el desarrollo de sistemas de búsqueda, tales como los sistemas recomendadores. 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 [Dublin, 2013] e IEEE LOM [IEEE LOM, 2013].

Por otra parte, el concepto Repositorios de Objetos de Aprendizaje 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 A., 2009].

Uno de los grandes retos en el área es propiciar el uso de los repositorios institucionales, 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, a partir de la calidad de los metadatos descriptivos. Aunque existen diferentes propuestas para la extracción automática de metadatos a partir de OA, 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 repositorios institucionales. 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.

En general, este trabajo aborda la problemática de la extracción automática de metadatos a partir de objetos de aprendizaje en repositorios institucionales, como un aspecto importante dentro de lo que es el acceso abierto y masivo a recursos digitales educativos. En la siguiente sección se presentarán los estándares más reconocidos y utilizados para metadatos, como son Dublin Core e IEEE LOM, y para objetos de aprendizaje, como SCORM; luego se revisará el fundamento y actualidad de los repositorios institucionales de acceso abierto de Argentina y Colombia, revisando las principales características de aquellos más representativos de cada país; en seguida se hará un enfoque en la extracción automática de metadatos en OA's, analizando cuatro sistemas: SAXEF, TWYS, MAGIC, Looking4LO, a la luz de los tres aspectos más importantes que se deben tener en cuenta en el momento de elegir o diseñar un sistema de extracción automática de metadatos: los tipos de archivo a procesar, los metadatos que se extraen y las técnicas y recursos utilizados para realizar la extracción; posteriormente se presentan comentarios y conclusiones acerca de esta comparación. Por último, se muestran las conclusiones y trabajos futuros a considerar en el área de extracción automática de metadatos.

Estándares para metadatos y objetos de aprendizaje

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 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 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.

Es posible encontrar dos tipos de metadatos en un OA: objetivos y subjetivos. Los primeros al estar relacionados al contenido del OA, sus valores 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 realizan el archivo del objeto en los repositorios.

La definición y adopción de estándares hace posible que los metadatos cumplan su función en mayor medida. Además, se hace necesario y recomendable para el almacenamiento y recuperación de los OA en los repositorios, ya que 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 [Dublin, 2013] e IEEE LOM [IEEE LOM, 2013], y para OA en general, SCORM [SCORM, 2013]; a continuación se detalla cada uno de estos estándares.

Dublin Core
DublinCore [Dublin, 2013] es una organización abierta, iniciada en 1995, que está abocada al desarrollo de estándares de metadatos interoperables. Este estándar se creó a partir de un taller de metadatos realizado precisamente en la ciudad de Dublin (Ohio), Estados Unidos. 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.

El estándar DCMI [DCMI, 2010], 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* (ver Figura 1). A su vez, estos se agrupan en 3 grandes categorías: **contenido**, **propiedad intelectual** e **instanciación.**

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). Así mismo, 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.

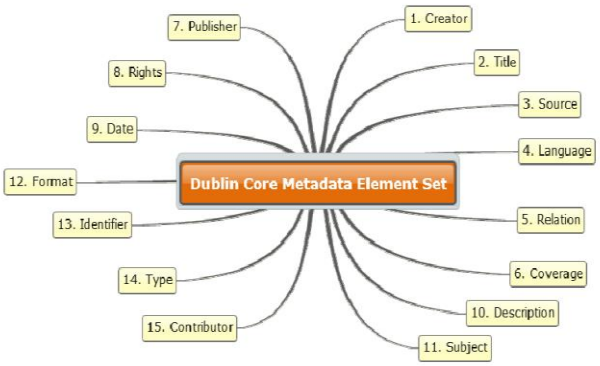

Fig. 1. Jerarquía completa de DCMI.

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 LOM, 2013]. 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 [IEEE LTSC, 2005]. 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 sub-categorías. 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.

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) [SCORM, 2013], 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.
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 OA la norma y uso de los metadatos, el empaquetamiento y secuenciamiento.

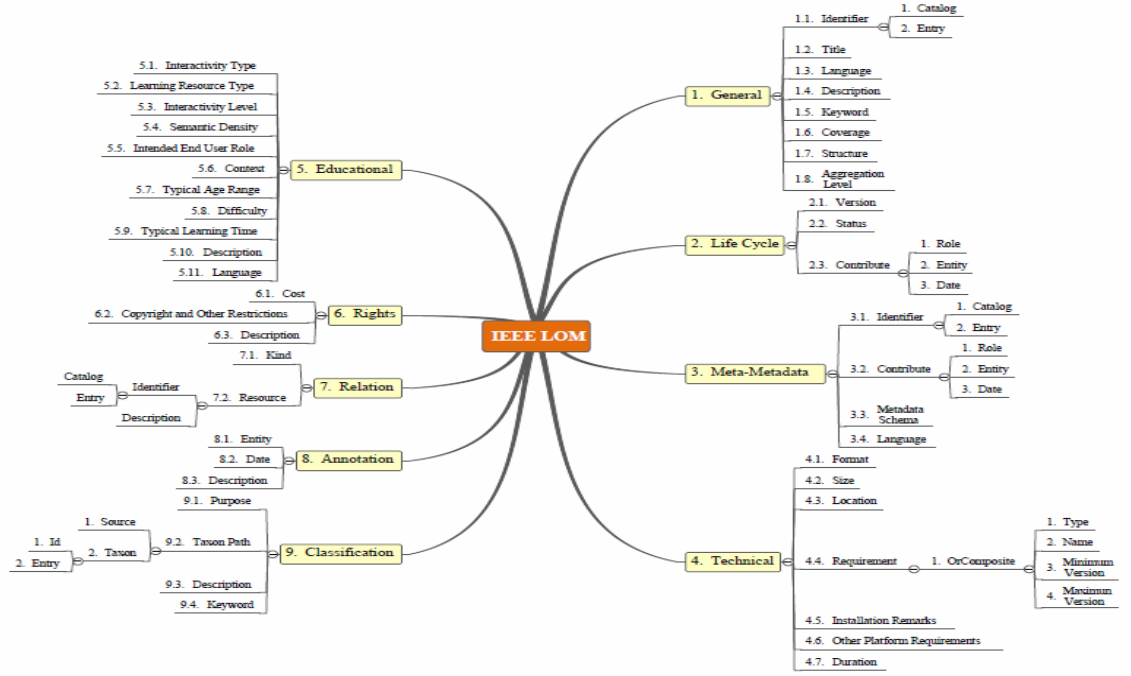

Fig. 2. Jerarquía completa de IEEE LOM.

Repositorios institucionales de 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 utilizado y reutilizado en el proceso de enseñanza-aprendizaje.

A continuación se realiza un diagnóstico sobre la situación actual de los repositorios institucionales de acceso abierto tanto en Argentina como en Colombia, mostrando,

para cada caso, las principales características de aquellos que se consideran más representativos.

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 [SNRD, 2013].

Así mismo, el 23 de Mayo de 2012, la Cámara de Diputados de la Nación Argentina dio media sanción al proyecto de Ley [Honorable Cámara, 2011] 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. Finalmente el 13 de noviembre de 2013 la cámara de senadores la aprueba convirtiéndola en ley.

La producción científica que será 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.

La interoperabilidad de los repositorios digitales que deberán crear las instituciones, será diseñada por el Sistema Nacional de Repositorios Digitales 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.

Repositorios institucionales en Colombia

En lo que respecta a lo existente en el área en Colombia [Recursos Educativos, 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 TIC, 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), se articulan los repositorios institucionales de las universidades colombianas encaminados a la promoción e instauración del acceso abierto como política educativa nacional. 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 TIC, 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 reciente Declaración de París de junio de 2012.

Análisis de repositorios institucionales de Argentina y Colombia

Teniendo en cuenta el contexto histórico de los repositorios institucionales de Argentina y Colombia presentado anteriormente, se realizó 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 esta revisión 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 [Ranking Web, 2013] 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 miden la visibilidad e impacto de los repositorios científicos".
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.

La revisión se realizó navegando a través de cada una de las páginas web de los repositorios, recopilando información de interés para el estudio de la importancia y significado de los metadatos dentro del proceso de búsqueda y recuperación de OA 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.

A continuación se describe brevemente cada uno de los repositorios analizados; del numeral 1) al 7) corresponden a repositorios institucionales de Argentina, mientras que del numeral 8) al 12) corresponden a repositorios institucionales de Colombia.

1) *SEDICI*: 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 4 a nivel Latinoamérica en el "Ranking Web de Repositorios del Mundo" [Ranking Web, 2013].
2) *Biblioteca Digital UNCuyo*: 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*: 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*: 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*: 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*: 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.
7) *BDU^2 Repositorios Institucionales*: 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.
8) *Repositorio Institucional UN*: es el repositorio de Acceso Abierto de la Universidad Nacional de Colombia, en el cual se pretende administrar, preservar y difundir toda 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 5 a nivel Latinoamérica en el "Ranking Web de Repositorios del Mundo" [Ranking Web, 2013].
9) *Repositorio Institucional EdocUR*: en este repositorio se da acceso a texto completo de los documentos producidos por la Universidad del Rosario en su función docente, investigativa y de extensión.
10) *Repositorio Institucional PUJ*: es el repositorio de la producción intelectual de la Pontificia Universidad Javeriana.

11) *BDCOL- Biblioteca Digital Colombiana*: 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*: este portal educativo, mencionado en la Tabla 4., 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 65 repositorios a nivel Latinoamérica dentro de 137 considerados en el "Ranking Web de Repositorios del Mundo" [Ranking Web, 2013], y dentro de los 10 primeros a nivel nacional, lo que da cuenta del gran trabajo y esfuerzo a nivel país realizado hasta el momento en lo que a desarrollo repositorios institucionales se refiere.
- La mayoría de los repositorios revisados utilizan plataforma DSpace (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) [DSpace, 2013] y estándar de metadatos DCMI, dando cuenta que este estándar es altamente usado a nivel de material educativo. 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 dejar de encontrar 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 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 mismo poder definir y asociar a cada categoría los respectivos metadatos que lo describen de la mejor manera.

Principales extractores de metadatos

Anteriormente se mencionó la importancia de la extracción automática de metadatos como una forma de garantizar la calidad de la información de los OA que se almacena en los repositorios y que será utilizada durante la ejecución de las búsquedas de material educativo, y de alguna manera, asistirá a los usuarios en la selección de OA's relevantes a sus preferencias y necesidades.

Esto, sumado a la estandarización de metadatos y el auge de los repositorios institucionales y del acceso abierto al conocimiento que se vislumbró y describió anteriormente, da como resultado el fundamento necesario para comprender la verdadera importancia del desarrollo de nuevos algoritmos para la extracción automática a partir de OA en repositorios institucionales.

Existen tres aspectos importantes que se deben tener en cuenta en el momento de elegir o diseñar un sistema de extracción automática de metadatos:

a) Los tipos de archivo a procesar (por ejemplo, html, txt, pdf, doc, etc.),
b) Los metadatos que extraen, 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.

En esta sección se presentan y analizan cuatro propuestas existentes para la extracción automática de metadatos:

1) SAFEX (System for Automatic eXtraction of E-learning object Features) [Alfano, 2006]: sistema creado por The Center on Communication Studies (Univ. Palermo, Italia), que automáticamente extrae indicadores didácticos de cualquier página Web.
2) TWYS (Tang Way Yuen System) [Wai Yuen, 2007]: este sistema fue desarrollado por Tang Way Yuen en su Tesis de Magister, dentro del departamento de Ciencias de la Computación en la Ciudad Universitaria de Hong Kong.
3) Looking4LO [Motz, 2009]: fue creado en el Instituto de Computación de la Facultad de Ingeniería (Universidad de la República, Uruguay). Looking4LO es un sistema genérico y flexible, capaz de extraer OA con sus respectivos metadatos de archivos XML y HTML, documentos Word, presentaciones Power Point, archivos PDF y paquetes SCORM.
4) MAGIC (Metadata Automated Generation for Instructional Content) [Li, 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.

En la Tabla 1 se presenta un cuadro comparativo de estas cuatro propuestas, a la luz de los tres aspectos relevantes en el diseño de un sistema de extracción automática de metadatos, mencionados anteriormente: tipos de archivos procesados, metadatos extraídos y técnicas y recursos utilizados.

Tabla 1. 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
SAFEX	HTML XHTML ASP PHP	No sigue al estándar LOM, 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

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.	Tokenizer POS Tagger Herramientas de procesamiento de lenguaje natural (TEXTRACT)

Con base en la comparación que se presenta en la Tabla 1, se pueden obtener los siguientes comentarios y conclusiones:

1. En cuanto a tipos de archivo, las cuatro propuestas analizadas pueden trabajar con páginas web HTML, debido, en primer lugar, a la gran cantidad de estos recursos de aprendizaje que se encuentra disponible en Internet y, en segundo lugar, porque las páginas web HTML tienen etiquetas que poseen más información acerca de su contenido. De manera particular, los archivos SCORM que son estructurados y tienen algunos campos de metadatos ya clasificados, sólo son tratados por Lookin4LO.
Algunos de los tipos de archivos no estructurados que no presentan etiquetas o metadatos, tales como TXT, PDF y PPT, son tratados por Looking4LO y MAGIC. Finalmente, MAGIC es el único sistema capaz de procesar otros tipos de objetos interesantes como lo son los archivos de video y audio. En general, de las cuatro propuestas revisadas MAGIC es la que contempla un amplio tipo de archivos más comúnmente utilizados.

2. En lo que respecta a los metadatos extraídos, tres de las cuatro propuestas revisadas extraen metadatos del estándar LOM, a excepción de SAXEF. TWYS es el sistema que mayor cantidad de metadatos extrae (tanto generales como educativos) siguiendo

el estándar LOM. MAGIC solamente permite obtener un subconjunto de los metadatos extraídos por TWYS, mientras que SAXEF extrae campos particulares definidos para un fin específico, que contienen información sobre la naturaleza del OA: títulos principales, títulos secundarios, teórico o práctico, sintético o analítico, tipo y nivel multimedia, nivel de complejidad y enlaces a otras tarjetas de identificación E-learning (EIC) con los mismos títulos o con títulos relacionados. Toda esta información se encuentra almacenada en la EIC del OA.

3. Cabe resaltar que algunas de las propuestas extraen metadatos educacionales y otros no, en parte, debido a que la extracción de estos metadatos no es trivial. 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 recomendadores en cuanto al ajuste de preferencias de acuerdo al tipo de usuario (profesor o estudiante). TWYS genera cuatro metadatos educacionales, a su vez, Looking4LO solamente extrae uno y MAGIC no extrae ninguno. En cuanto a SAXEF, en la EIC se discrimina cuando una página web es teórica o práctica y sintética o analítica, además provee un nivel multimedia.

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, mientras que otros metadatos requerirán reglas o métodos no triviales para determinar su valor. Además se observa que los sistemas Looking4LO y MAGIC, utilizan como apoyo herramientas de procesamiento de lenguaje natural (NLP por sus siglas en inglés).

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) No siguen un estándar de metadatos, lo que puede conducir a la diversidad y baja calidad en la información descriptiva asociada a los OA.
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.

Conclusiones y trabajos futuros

Este trabajo permitió evidenciar que cada vez toman mayor fuerza e importancia los repositorios institucionales de acceso abierto, que se apoyan 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 comunicad educativa atendiendo los distintos niveles de usuarios generalmente no técnicos, dentro de estas últimas dificultades está el entendimiento de los metadatos como descriptores apropiados para encontrar, gestionar, reusar y almacenar OA en forma efectiva.

Teniendo en cuenta el análisis que se realizó de los cuatro sistemas extractores de metadatos, es evidente que en lo que respecta a esta tecnología es mucho el camino que queda por recorrer, tanto en aquellos tipos de archivos conocidos y comunes como documentos de texto y PDF, como en otros tales como los archivos tipo música e imágenes que tienen un mayor grado de dificultad. Esto es así 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. Por otra parte, la generación semi-automática 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, incompletitud, inconsistencias y discrepancias.

Así mismo, después de haber hecho el análisis de los metadatos, estándares y repositorios en su conjunto, se observa que ninguna de las propuestas hace uso integrado de los estándares DublinCore e IEEE LOM, y que, por tal motivo, no es posible garantizar la interoperabilidad entre los repositorios que usan distintos estándares y así ampliar los horizontes de acceso a la información. Por lo tanto, en la generación de nuevos algoritmos y herramientas para la extracción automática de metadatos, sería interesante tener en cuenta esta perspectiva de trabajo, como una forma de aprovechar los grandes esfuerzos que se han hecho hasta el momento para brindar la posibilidad de estandarizar y unificar en mayor grado las diversas maneras que existen de almacenar OA.

En unión con lo anterior, otro punto de investigación puede enfocarse en establecer métodos de ponderación y evaluación de 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 mejora y retroalimentación de los algoritmos de extracción automática de metadatos, y por consiguiente, mejoraría los resultados de las búsquedas que se realizan.

También 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 en 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

1. Alfano, M., Lenzitti, B.,Visalli, N. (2007). SAXEF: A System for Automatic eX-traction of learning object Features. Journal of e-Learning and Knowledge Society, vol. 3, (2), 83-92.
2. Casali A., Gerling V., Deco C. y Bender C. (2009). Un Sistema inteligente para asistir la búsqueda personalizada de objetos de aprendizaje. Recuperado a partir de http://rephip.unr.edu.ar/bitstream/handle/2133/2816/Gerling.pdf?sequence=1
3. DCMI. (2010). Metadata Basics. Página Web, . Recuperado Septiembre 19, 2013, a partir de http://dublincore.org/metadata-basics/
4. DSpace. (2013). Página Web, . Recuperado Octubre 23, 2013, a partir de http://www.dspace.org/
5. Dublin Core Metadata Initiative. (2013). Página Web, . Recuperado Septiembre 19, 2013, a partir de http://www.dublincore.org
6. Honorable Cámara de Diputados de la Nación, Proyecto de Ley. (2011). Creación de Repositorios Digitales Institucionales de Acceso Abierto, Propios o Compartidos. Página Web, . Recuperado Septiembre 13, 2013, a partir de http://www1.hcdn.gov.ar/proyxml/expediente.asp?fundamentos=si&numexp=1927-D-2011
7. IEEE LOM. (2013). Página Web, . Recuperado Septiembre 19, 2013, a partir de http://www.ieee.org
8. IEEE LTSC. (2005). Position Statement on 1484.12.1 - 2002 Learning Object Metadata (LOM) Standard Maintenance/Revision. LOM Standard Maintenance/Revision. Página Web, . Recuperado Septiembre 19, 2013, a partir de http://ltsc.ieee.org/news/20021210-LOM.html
9. Li, Y., Dorai, C., Farrell, R. (2005). Creating MAGIC: system for generating learning object metadata for instructional content, in MULTIMEDIA '05: Proceedings of the 13th annual ACM international conference on Multimedia, pp. 367-370, New York, NY, USA.
10. Motz, R., Badell, C., Barrosa, M., Sum, R., Díaz, G., Castro, M. (2009). LooKIng4LO: Sistema Informático para la Extracción Automática de Objetos de Aprendizaje: Caso de Estudio. IEEE-RITA(2009) 223-229.
11. Polsani, P. R. (2003). Use and Abuse of Reusable Learning Objects. Journal of Digital Information, 3(4). Recuperado a partir de journals.tdl.org/jodi/article/viewArticle/89
12. Ranking Web de Repositorios. (2013). Consejo Superior de Investigaciones Científicas, CSIC. Página Web, . Recuperado Octubre 21, 2013, a partir de http://repositories.webometrics.info/es
13. Recursos Educativos Digitales Abiertos. (2012). Colombia. Ministerio de Educación Nacional. Recuperado a partir de http://www.colombiaaprende.edu.co/html/home/1592/articles-313597_reda.pdf
14. SCORM. (2013). Página Web, . Recuperado Octubre 10, 2013, a partir de http://www.adlnet.org/scorm/

15. Sistema Nacional de Repositorios Digitales. (2013). Página Web, . Recuperado Septiembre 13, 2013, a partir de http://repositorios.mincyt.gob.ar/
16. Wai Yuen, T. (2007). Automatic Extraction of Learning Object Metadata (LOM) from HTML Web Pages, Master of philosophy, City University of Hong Kong, May 2007
17. Wiley, D. (2002). Connecting Learning Objects to Instructional Design Theory: A definition, a metaphor, and a taxonomy, in *Instructional Use of Learning Objects*, edited by D. A.Wiley, Ed. Association for Instructional Technology.