*T. Pire, B. Espinasse, A. Casali, C. Deco, (2011), « Automatic Extraction of Learning Objects Metadata for Recommendation: a comparative study ». InfoEdu 2011, Convergent Technologies, Integration and Independence Conference, Habana, Cuba, February 7-11, 2011.*

# Extracción Automática de Metadatos de Objetos de Aprendizaje: un estudio comparativo

**Taihú Pire[1], Claudia Deco[1,2] , Ana Casali [1,3] y Bernard Espinase[4]**

[1] Facultad de Ciencias Exactas, Ingeniería y Agrimensura, Universidad Nacional de Rosario
taihup@gmail.com, acasali@fceia.unr.edu.ar, deco@fceia.unr.edu.ar
[2] Facultad de Química e Ingeniería Rosario, Universidad Católica Argentina, Rosario, Argentina
[3]Centro Internacional Franco Argentino de Ciencias de la Información y de Sistemas - CIFASIS, Rosario, Argentina.
[4] Laboratoire de Science de l´Information et des Systèmes – LSIS, Marseille, France,
bernard.espinasse@lsis.org

## Resumen

En la última década, Internet se ha utilizado, entre otras cosas, como fuente de información educativa. Para ayudar en el almacenamiento, clasificación y reutilización de recursos educativos aparece el concepto de Objetos de Aprendizaje (OA), con el fin de clasificar el material educativo para proporcionar unidades modulares de aprendizaje con metadatos, y para mejorar el acceso y la reutilización de los mismos. En este trabajo se analizan, por un lado, la importancia de los metadatos de los objetos de aprendizaje con el fin de poder utilizarlos en un sistema de recomendación automática personalizada. Por otro lado, se explora el estado del arte de las técnicas de extracción automática de metadatos, y se analizan y comparan diferentes sistemas de extracción. Por último, se presentan algunas conclusiones sobre posibles líneas de investigación para abordar el problema de la falta de información de metadatos en los objetos de aprendizaje.

**Palabras claves**: Objetos de aprendizaje, metadatos educacionales, extracción de información, estándar LOM.

## Abstract

In the last decade, Internet is used, among others things, as an educational information source. To help in storage, classification and reuse of educational resources appears the concept of Learning Objects (LO) in order to classify educational material, to provide modular units of learning with metadata, and to improve the access and reuse of them. In this work we analyze, on the one hand, the importance of metadata in Learning Objects in order to obtain a personalized recommendation. On the other hand, exploring the state of the art of automatic metadata extraction, we analyze different software systems and we make a comparison of these systems. Finally, we make some conclusions about several lines of possible research work to address the problem of lack of metadata information in LOs.

**KeyWords**: Learning Objects, educational metadata, information extraction, LOM standard.

## Introducción

Hoy en día, la Web es una de las más importantes fuentes de recursos educacionales, donde los estudiantes y profesores tienen una gran cantidad de información a su disposición. Para el proceso de recuperación de información, las personas usan motores de búsqueda, como Google o Yahoo, los cuales, desafortunadamente en la mayoría de los

casos, no retornan la información deseada o devuelven una gran cantidad enlaces.
Para ayudar al almacenamiento, clasificación y reutilización de los recursos educacionales, surge el concepto de Objetos de Aprendizaje (OA). Un OA es "cualquier recurso digital que puede ser reutilizado para la enseñanza" [1]. Éstos pueden ser usados por un estudiante que quiere aprender un determinado tema o por un profesor que quiere preparar algún material para su clase. Los OAs son descritos por metadatos que frecuentemente siguen el estándar LOM[1]. Los usuarios pueden recuperar OAs por medio de búsquedas en repositorios Web. Algunos ejemplos de estos repositorios son: FLOR[2], Ariadne[3], y OER Commons[4].

Dada una consulta sobre un tema específico, diferentes usuarios obtienen como resultado la misma lista de objetos. Generalmente el usuario concentra su atención en los primeros resultados, los cuales no siempre le sirven si la búsqueda se hace considerando sólo las palabras claves. Esto sucede porque los usuarios tienen distintas características y preferencias, que deben ser consideradas al momento de la búsqueda. Los sistemas recomendadores logran resolver este tipo de problemas dado que pueden seleccionar el material que es más apropiado para las necesidades y características de los usuarios.

En la siguiente sección analizaremos la importancia de los metadatos en los objetos de aprendizaje con el fin de obtener una recomendación personalizada. Luego, haremos un enfoque en la extracción automática de OAs, analizando cuatro sistemas de extracción de metadatos: SAXEF, TWYS, Looking4LO y MAGIC. Posteriormente, haremos una comparación entre dichos sistemas tomando algunas características específicas de los mismos. Finalmente, analizaremos este estudio comparativo y propondremos algunas posibles líneas de investigación para aminorar la falta de metadatos en los OAs.

## Importancia de los metadatos

El desarrollo de un sistema recomedador parece ser un buen camino para asistir a los usuarios en la selección de OAs relevantes a sus preferencias y necesidades. En los últimos años, la comunidad de Inteligencia Artificial ha realizado un gran trabajo sobre sistemas recomendadores [2]. Esta clase de sistemas pueden ayudar a las personas a encontrar lo que realmente desean, en Internet específicamente, tomando en cuenta sus preferencias personales.
Existen algunos enfoques para la personalización de los resultados, que toman en cuenta el perfil del usuario, incluyendo sus características y preferencias ([3], [4]). Estos sistemas utilizan metadatos con descripciones semánticas de los objetos. Particularmente en [5] se propone un sistema recomendador para el dominio educacional. La arquitectura presentada está basada en un sistema multiagente, el cual permite trabajar de una manera flexible con información distribuida en repositorios. Los resultados obtenidos son mostrados en un ranking de OAs.

Sin embargo, un problema encontrado es la falta de información en muchos de los metadatos educacionales de los objetos de aprendizaje ubicados en los repositorios y en la calidad de los metadatos [6]. En relación a la búsqueda personalizada de cursos sobre educación a distancia, Sonntag analiza en [7] la importancia de los metadatos para los objetos de aprendizaje, exponiendo que estos recursos educativos pueden ser frecuentemente reutilizados y posiblemente en diferentes contextos. Algunos de los problemas que destaca son la escasez de metadatos y la diversidad de estándares. Por otro lado, preparar objetos de aprendizaje con metadatos adecuados es laborioso, y a veces lleva a una falta de calidad en los metadatos.
Un análisis de la información en los metadatos de OAs en algunos repositorios se muestra en

[1] http://www.ieee.org
[2] http://www.laclo.org
[3] http://www.ariadne-eu.org
[4] http://www.oercommons.org

[8]. El trabajo considera los repositorios: FLOR, OER Commons y Ariadne. Se observa que tanto FLOR como Ariadne soportan la realización de consultas en una federación de repositorios. Dicho autor señala la escasez de información en los metadatos educacionales, los cuales son una subcategoría de los metadatos LOM. La mayoría de los objetos en OER Commons sólo poseen metadatos que especifican el rango de edad estimado del usuario al cual va dirigido el objeto y el idioma, pero no tiene en cuenta ni el tiempo estimado de aprendizaje que necesitará el usuario para aprender el tema ni el nivel de interactividad que requiere el objeto por parte del usuario. Los recursos incluyen algunos otros campos de LOM como lo son el tipo de contenido y el nivel académico. A su vez, los OAs en FLOR tienen metadatos educacionales como tipo de recurso de aprendizaje, rol del usuario destino y contexto. Sus objetos carecen de los metadatos tipo de interactividad requerido y dificultad, ambos importantes para la realización de una recomendación personalizada. Por otro lado, Ariadne es el más completo de todos los repositorios teniendo OAs con metadatos de casi todos los campos educacionales. Considerando la calidad de los metadatos, existen distintos problemas que surgen desde el análisis realizado en [8]. Por ejemplo, el valor del metadato idioma de un OA, fue establecido usando únicamente el título, pero el cuerpo del objeto estaba en otro idioma. Además un documento fue clasificado como texto en el metadato tipo de recurso de aprendizaje, pero éste era un código de un programa de computación y por lo tanto debería ser clasificado mejor como Ejercicio o Ejemplo.

Debido a la importancia de los metadatos para la recuperación personalizada de OAs y a la falta de calidad de información de estos metadatos, el desarrollo de sistemas para la extracción automática de metadatos parece ser un paso muy importante hacia la resolución de este problema. En esta dirección, el objetivo de este trabajo es analizar distintos métodos de extracción automática de información que son capaces de completar campos de metadatos y detectar nuevas líneas de investigación en el área. En la siguiente sección, nos enfocaremos en su extracción automática.

## Extracción automática de metadatos de OAs

Hasta el día de hoy, no se ha trabajado demasiado en la extracción automática de metadatos. Cada herramienta para la extracción de metadatos tiene sus propios objetivos, arquitectura y usa distintas técnicas. Esta sección presenta y analiza cuatro sistemas dedicados a la extracción automática de metadatos de objetos de aprendizaje: SAXEF, TWYS, Looking4LO, y MAGIC.

### El Sistema SAFEX

*SAXEF* (System for Automatic eXtraction of E-learning object Features) ([9], [10]) es un sistema creado por The Center on Communication Studies (Univ. Palermo, Italia), que automáticamente extrae indicadores didácticos de cualquier página Web. Produce una tarjeta de identificación de E-learning (EIC, por sus siglas en inglés) que permite a los profesores evaluar fácilmente cuándo una página es de su interés.

El sistema SAXEF ha sido pensado para la extracción de atributos texto/multimedia desde cada página Web (considerada como un objeto de aprendizaje) o un grupo de páginas Web (que representan un curso completo). En la práctica, dado un curso o simplemente un objeto de aprendizaje, SAXEF produce la EIC. Las EICs son organizadas en una base de datos y son mostradas a través de una interfaz gráfica indicando tema principal y sus conexiones. SAXEF elimina las palabras comunes de un texto para quedarse con las relevantes (palabras claves y títulos). Para llevar a cabo esta tarea, utiliza un archivo de texto que contiene artículos, preposiciones, pronombres, verbos comunes, etc. Más aun, identifica palabras relevantes que se encuentran dentro de etiquetas como <title> y <meta> en un archivo HTML. Luego, a cada palabra se le asigna un peso, que es utilizado

para determinar el título principal y los secundarios. En la práctica, el peso es un puntaje que se obtiene dependiendo de cuándo y dónde aparece la palabra en el texto. Para llevar a cabo el análisis multimedia, SAXEF procesa las áreas textuales y multimedias. El área textual está determinada por el producto entre el número de caracteres de la página Web y el área ocupada por cada carácter. El área multimedia está determinada por la suma de las áreas de los objetos multimedia presentes en la página Web. En particular, son considerados el tamaño (en pixeles) de las imágenes, videos y animaciones. Además, si existe un archivo de audio, se considera su tamaño (en bits) dividido por 16 bits (tamaño de muestreo). Ahora, asumiendo que el área de la página Web es la suma del área textual y el área multimedia, el cociente entre el área textual y el área total resultará en el índice analítico (expresado como un porcentaje). Al mismo tiempo, el cociente entre el área multimedia y el área total es el índice sintético complementario. Para obtener los tipos multimedia y el nivel multimedia (expresado como un porcentaje), SAXEF considera cuántos objetos multimedia aparecen en la página Web y el área ocupada por los mismos.

**El Sistema TWYS**

*TWYS*[5] [11] es un sistema desarrollado por Tang Way Yuen, dentro del departamento de Ciencias de la Computación en la Ciudad Universitaria de Hong Kong. TWYS es capaz de extraer metadatos de objetos de aprendizajes desde páginas HTML. Está basado en el estándar IEEE LOM.

TWYS pretende ayudar a los usuarios a adquirir fácilmente recursos de aprendizaje relevantes haciendo que los motores de búsqueda adopten y soporten el estándar IEEE LOM por medio de la extracción automática desde páginas webs. Parte de la información encontrada en las páginas Web puede ser traducida a algunos elementos LOM, mientras que otras requerirán reglas o métodos no triviales para determinar el valor de los elementos LOM. Tang en [11] propone dos métodos:

- *Mapeo directo*, refiere a información HTML que puede ser interpretada a elementos LOM directamente. Algunas técnicas son *StopWord*, *Term frecuency weighting* (TF-IDF) y Ontologías.
- *Reglas basadas en heurísticas*, son aplicadas a otros valores que no pueden ser obtenidos desde la información propia de una página Web. Se utilizan dos técnicas: (1) Verificación de la Existencia de Etiquetas HTML -*Check Existence of HTML Tag*-, de esta manera se puede saber cuándo cierta etiqueta HTML existe en la página; y (2) Conteo Estático de de HTML y Contenido -*Static Counting of HTML and Content*-, cuenta la cantidad de veces que aparecen ciertas etiquetas HTML y ciertas palabras en una página Web.

Al comienzo, TWYS aplica un crawler para colectar las páginas y luego almacenarlas en una base de datos de archivos HTML. Luego, separa el contenido HTML de las cabeceras y de las etiquetas HTML por medio de un parser. Este último elimina información inservible del contenido de la página a través de la base de datos de *StopWords*, generando un archivo pre-procesado. Luego, el parser utiliza el vocabulario predefinido por la Ontología y verifica si alguna de las palabras contenidas en el vocabulario están presentes en el archivo pre-procesado. Si ninguna de las palabras del vocabulario existe en el archivo, se marca como no relevante al dominio de interés. El parser por lo tanto descartará la página Web de la base de datos, manteniendo así la relevancia de los documentos del repositorio. Si el archivo coincide con el dominio de interés, el parser tomará todos los sub-enlaces de las páginas y los enviará nuevamente al crawler para aumentar la colección. Finalmente, TWYS extrae y genera

[5] Adoptamos la breviacion "TWYS" (Tang Way Yuen System) para el sistema creado por Tang Way Yuen en su Tesis de Master.

registros LOM por medio de métodos y heurísticas.
TWYS puede obtener de la cabecera HTML y las etiquetas los campos del estándar LOM Entry, Location, Title, Language, Entity, Date, Format y Size. Utiliza los métodos StopWors y TF-IDF para la obtención de los atributos Description y Keyword del contenido de la página. Para producir los campos LOM Purpose, ID, Entry, Description y Keywords, TWYS hace uso de la ontología para el contenido. Como ya mencionamos, TWYS usa algunas reglas basadas en heurísticas para lograr los campos educacionales del estándar LOM. Para obtener el campo Interactivity Type, verifica la existencia de las etiquetas HTML <Form action=> e <Input>. Para conseguir el Interactivity Level, realiza una cuenta de las etiquetas <Input> y <action=>. Para extraer el campo Semantic Density, suma la cantidad de palabras y datos multimedia. Finalmente, para conseguir el atributo Difficulty, cuenta el número de palabras distintas de la página.

**El Sistema Looking4LO**

*Looking4LO* [12] 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 objetos de aprendizaje con su respectivos metadatos de archivos XML y HTML, documentos Word, diapositivas Power Point y archivos PDF, como también de paquetes SCORM. Looking4LO es un prototipo que puede extraer OAs que cubren una cierta temática (por ejemplo: matemática, lógica, historia, gastronomía, etc.) y puede automáticamente proveerlos con metadatos. Para la implementación de la extracción de OAs, Looking4LO utiliza la plataforma GATE[6] (General Architecture for Text Engineering). En GATE, el contenido de un documento es modelado como un conjunto de anotaciones, es decir información adicional acerca de fragmento particular del contenido del documento. Las anotaciones son creadas y modificadas por diferentes recursos de procesamiento. Para poder generar estas anotaciones, un recurso de procesamiento toma como entrada el documento a procesar (representación interna de GATE) y quizás necesite otros recursos previamente creados, como por ejemplo, ontologías u otras anotaciones. La salida es un documento enriquecido con nuevas anotaciones y/o modificaciones de las anteriores. El motor de extracción de Looking4LO está integrado por cinco recursos de procesamiento: *Tokenizer, Sentence Splitter, POS Tagger, Gazetteer y Transducer.*
Las entradas del sistema son recibidas por un controlador que identifica los tipos de documentos a ser analizados y luego los envía con el resto de los parámetros de entrada a la respectiva unidad de procesamiento, llamadas *Wrapper*. Cada wrapper maneja un tipo diferente de documento. Dentro del wrapper un motor de extracción es el responsable de analizar el contenido del documento. Para llevar a cabo esta tarea el proceso de extracción sigue los siguientes pasos. Primero, el Tokenizer identifica las marcas (tokens) en el documento. Luego, el Sentence Splitter delimita las oraciones. Después, el POS Tagger determina la categoría gramatical de cada token. Luego, el Gazetter identifica los conceptos en el documento que pertenecen al dominio de la ontología. Esta ontología modela el área temática, eso significa que el Gazetter permite detectar qué fragmentos del documento pertenecen a algún concepto relevante; el Transducer ejecuta un conjunto de reglas contextuales para identificar las ocurrencias de los OAs. Finalmente, los OAs resultantes son empaquetados por el Packager en el formato deseado. Durante el proceso de recuperación, los metadatos extraídos para cada OA son Author, Read time, Interactivity Level y detecta si el OA tiene imágenes o no.

[6] GATE (http://gate.ac.uk/) es una herramienta que permite integrar componentes del Procesamiento del lenguaje Naturalntegrating para la construcción de diferentes aplicaciones.

**El Sistema MAGIC**

*MAGIC* (Metadata Automated Generation for Instructional Content) [13] es un sistema desarrollado en el centro de investigacion IBM Watson, que automáticamente identifica, segmenta y genera metadatos críticos de acuerdo al estándar SCORM para contenidos educacionales. Puede procesar tanto archivos de texto como multimedia.
MAGIC apunta a extraer metadatos SCORM de diferentes tipos de archivos (por ejemplo, video, audio, texto, etc.) aplicando un conjunto de métodos de extracción de información. Su principal objetivo es asistir a profesores y a desarrolladores de cursos con la adopción de SCORM y permitir la reutilización de información. Aquí, sólo nos enfocaremos en su método de extracción de metadatos para texto.
MAGIC lleva a acabo la tarea de extraer metadatos de textos con una herramienta llamada TEXTRACT [14]. Este sistema es un componente integrado por una importante infraestructura para el análisis y procesamiento de documentos, que hace uso de filtros lingüísticos. TEXTRACT fue diseñado para realizar una gran variedad de funciones de extracción lingüística. Algunas de estas funciones son relativamente inmediatas, por ejemplo, tokenización, búsqueda léxica y análisis morfológico. Otras son más complejas, como la extracción de terminología técnica y la extracción de las frases más importantes a lo largo de una colección de documentos.
El sistema MAGIC sigue los pasos que se detallan a continuación. Primero, el documento de texto es tokenizado. Para esta tarea, usa *Frost* (un componente del producto IBM LanguageWare™). Luego, cada palabra es etiquetada con su semiologia (adjetivo, sustantivo o verbo). Luego, se aplican módulos analíticos (no explicitados en la información provista públicamente por los autores del sistema) para la extracción de los siguientes metadatos: Title, Keyword, Entity y Description. Finalmente, se genera un archivo de metadatos LOM agregando cada metadato en formato XML utilizando la información extraída.

**Otros Sistemas**

Además de los sistemas presentados, hemos considerado otros dos sistemas de extracción: TextWise y la API Alchemy.
*TextWise*[7] es una plataforma que permite crear el ADN semántico de documentos, revelando el significado fundamental del texto. Semantic Signatures® trata el contenido para dar a luz el verdadero significado del texto y crear una única firma para cada documento procesado.
*Alchemy API*[8] es un producto de Orchestr8, una compañía enfocada en el etiquetado semántico y soluciones para el tratamiento de textos. Usa tecnologías basadas en estadísticas para el procesamiento del lenguaje natural y algoritmos de aprendizaje automatizado para el análisis de contenido y extracción de metadatos semánticos, por ejemplo información acerca de personas, lugares, compañías, títulos, idiomas entre otros. La finalidad de la API es proveer un análisis de contenido para páginas Web accedidas a través de Internet y escanear imágenes.
Estos dos sistemas, TextWise y Alchemy API, son interesantes ya que son open source y están disponibles para extraer algunos tipos de metadatos. Pero como no pudimos obtener detalles referentes a qué técnicas particulares utilizan, no fueron incluidos en la comparación con los cuatro sistemas considerados.

## Estudio comparativo sobre los sistemas de extracción analizados

Esta sección compara los sistemas SAXEF, TWYS, Looking4LO y MAGIC, tomando como puntos de comparación los siguientes: tipos de archivo tratados, metadatos extraídos y técnicas utilizadas para la extracción automática. Se pueden distinguir dos características importantes en la extracción de metadatos que sigan el estándar LOM: la

[7] http://textwise.com/
[8] http://www.alchemyapi.com

naturaleza de los tipos de archivo tratados por la extracción y la naturaleza de los metadatos extraídos.

*Tipos de Archivos*: una de las principales características que deben ser consideradas al momento de trabajar con un sistema de extracción de metadatos es que éste debe poder procesar distintos tipos de archivo. Ejemplos de éstos son: HTML, TXT, PPT, PDF, documentos Word y archivos de video. HTML es un texto estructurado; esto significa que las páginas Web HTML tienen etiquetas que definen su título, idioma, fecha y tamaño entre otros. Por lo tanto, éstos tienen más información acerca de su contenido que los otros tipos mencionados anteriormente.

*Metadatos extraídos*: es uno de los puntos más importantes a ser considerados. Algunos de los sistemas presentados aquí extraen campos pertenecientes al estándar LOM mientras que otros extraen campos específicos definidos para un determinado fin.

*Recursos utilizados en el proceso de extracción*: para el proceso de extracción automática de metadatos cada sistema puede usar distintas recursos de procesamiento, tokenizadores, POS taggers, ontologías, etc.

En la siguientes subsecciones, compararemos los cuatro sistemas, dedicados a la extracción automática de metadatos LOM, de acuerdo a estos tres puntos de comparación.

**Comparación de Tipos de Archivo**

La mayoría de los sistemas de extracción pueden manipular archivos HTML. En particular, los seis sistemas presentados aquí pueden tratar páginas Web HTML. También, los archivos SCORM son estructurados y tienen algunos campos de metadatos ya clasificados. Estos tipos de archivos son tratados por Looking4LO.
Algunos de los tipos de archivos no estructurados (archivos TXT, PDF, y PPT), es decir que no presentan etiquetas o metadatos, son tratados por Looking4LO, MAGIC, TextWise y Alchemy API. En particular, si un texto plano tiene una estructura específica, esto es, si sigue ciertas reglas o formato (por ejemplo, si tenemos apuntes compuestos por una definición, seguida por un ejemplo y finalmente por un ejercicio) los extractores pueden tomar ventaja de esta situación y hacer una mejor recuperación de metadatos.
Otros tipos interesantes de manejar son los archivos de video y audio (AVI, MPG, MP3, MP4, WMA). MAGIC es el único de los sistemas presentados aquí que puede manipular esta clase de formatos (otros extractores que pueden llevar a cabo esta tarea son Anvil, Elan, EMARaLDA, TASX, MacVisTA; en [14] se puede encontrar una comparación entre ellos).

La Tabla 1 muestra los tipos de archivos tratados por cada uno de los cuatro sistemas considerados. Por un lado, considerando los tipos de archivos, todos los sistemas pueden trabajar con archivos HTML, ya que son más fáciles de manejar que otros tipos de archivos. Hay dos sistemas que pueden manipular únicamente páginas Web: SAXEF y TWYS. A su vez, Looking4LO, MAGIC y TextWise pueden manipular diversos tipos de archivos (HTML, XHTML, ASP, PHP, PPT). Por otro lado, respecto a los metadatos extraídos, TWYS es el sistema que mayor cantidad de metadatos extrae (tanto generales como educacionales). Sin embargo, dicho sistema sólo puede trabajar con archivos HTML.

**Comparación de Metadatos Extraídos**

El poder de extracción de metadatos es uno de los puntos más importantes a ser considerados. Algunos de los sistemas presentados extraen campos del estándar LOM, mientras que otros extraen campos particulares, definidos para un determinado fin. SAXEF no sigue al estándar LOM, ya que produce una tarjeta de identificación de E-learning (EIC) con la siguiente información sobre la naturaleza del curso/objeto: (i) títulos principales; (ii) títulos secundarios; (iii) teórico o práctico; (iv)

sintético o analítico; (vi) tipo multimedia y nivel multimedia; (vii) nivel de complejidad; (viii) enlaces a otras EICs con los mismos títulos; y (ix) enlaces a otras tarjetas con títulos relacionados.

Tabla 1: Comparación de Sistemas
(Tipos de Archivos y Metadatos extraídos)

| **Sistema** | **Tipo de Archivo** | **Metadata extraída** |
|---|---|---|
| SAXEF | HTML<br>XHTML<br>ASP<br>PHP | - Secondary topics<br>- Theoretical or Practical<br>- Synthetic or Analytical<br>- Media types and multimediality level<br>- Complexity level<br>- Links to other EICs with same topics<br>- Links to other EICs with related topics |
| TWYS | HTML | - Entry (LOM 1.1.2)<br>- Location (LOM 4.3)<br>- Title (LOM 1.2)<br>- Language (LOM 1.3)<br>- Entity (LOM 2.3.2)<br>- Date (LOM 2.3.3)<br>- Format (LOM 4.1)<br>- Size (LOM 4.2)<br>- Description (LOM 1.4)<br>- Keyword (LOM 1.5)<br>- Purpose (LOM 9.1)<br>- ID (LOM 9.2.2.1)<br>- Entry (LOM 9.2.2.2)<br>- Description (LOM 9.3)<br>- Keywords (LOM 9.4)<br>- Interactivity Type (LOM 5.1)<br>- Interactivity Level (LOM 5.3)<br>- Semantic Density (LOM 5.4)<br>- Difficulty (LOM 5.8) |
| Looking4LO | HTML<br>PDF<br>TXT<br>PPT<br>SCORM | - Author (LOM 2.3.2)<br>- Reading time<br>- Has image<br>- Interactivity level (LOM 5.3) |
| MAGIC | HTML<br>PDF | - Title (LOM 1.2)<br>- Keyword (LOM 1.5)<br>- Entity (LOM 2.3.2)<br>- Description (LOM 1.4) |
| TextWise | HTML<br>PDF<br>TXT<br>WORD | Información no disponible |
| Alchemy API | HTML<br>TXT | Información no disponible |

A diferencia de SAXEF, TWYS adopta el estándar LOM. Este último extrae los siguientes campos: Entry (LOM 1.1.2), Location (LOM 4.3), Title (LOM 1.2), Language LOM 1.3), Entity (LOM 2.3.2), Date (LOM 2.3.3) Format (LOM 4.1), Size (LOM 4.2), Description (LOM 1.4), Keyword (LOM 1.5), Purpose (LOM 9.1), ID (LOM 9.2.2.1), Entry (LOM 9.2.2.2), Description (LOM 9.3), Keywords (LOM 9.4), Interactivity Type (LOM 5.1), Interactivity Level (LOM 5.3), Semantic Density (LOM 5.4) y Difficulty (LOM 5.8). El sistema TWYS extrae la mayor cantidad de campos siguiendo el estándar LOM.

MAGIC solamente permite obtener un subconjunto de los metadatos extraídos por TWYS: Title (LOM 1.2), Keyword (LOM 1.5), Entity (LOM 2.3.2) y Description (LOM 1.4). Por otro lado, Looking4LO extrae algunos campos LOM y otros campos de interés: Reading time, Image, Author y Interactivity level. Los últimos dos metadatos pertenecen al estándar LOM (LOM 2.3.2 y LOM 5.3).

Un punto interesante a tener en cuenta, es que algunos de estos sistemas extraen metadatos educacionales y otros no. Las EICs de SAXEF discriminan cuándo una página Web es teórica o práctica y sintética o analítica. Además, provee un nivel multimedia para las mismas.

TWYS genera cuatro metadatos educacionales: Interactivity Type, Interactivity Level, Semantic Density y Difficulty. La extracción de estos metadatos no es trivial. Para esto TWYS usa reglas basadas en heurísticas. A su vez, Looking4LO solamente extrae Interactivity level, y MAGIC no extrae ninguno campo. La Tabla 1 muestra los diferentes metadatos extraídos por los sistemas de extracción automática considerados.

**Comparación de Técnicas Utilizadas**

La Tabla 2 ilustra las diferentes técnicas utilizadas por los sistemas considerados para lograr la extracción automática.

Tabla 2: Comparación de Sistemas (Recursos de procesamiento y herramientas NLP utilizadas)

| Sistema | Recurso de Procesamiento | Herramienta NLP utilizada |
|---|---|---|
| SAXEF | Stop Words<br>Direct Mapping rules<br>Heuristic Mapping rules<br>Statistical Measure | Ninguna |
| TWYS | Ontología<br>Stop Words<br>TF-IDF<br>HTML Parser<br>Direct Mapping rules<br>Heuristic Mapping rules | Ninguna |
| Looking4LO | Ontología<br>Tokenizer<br>Sentences Splitter<br>POS Tagger<br>Gazetteer<br>Transducer | GATE |
| MAGIC | Tokenizer<br>POS Tagger | TEXTRACT |

En esta tabla se detalla para cada sistema, los recursos de procesamiento y las herramientas de procesamiento de lenguaje natural (NLP, por su sigla en inglés) utilizadas para la extracción.

## Conclusión

En este trabajo se presentó la importancia de los metadatos de los objetos de aprendizaje para dar una recomendación personalizada. El análisis de los distintos repositorios ha mostrado que hay una falta de calidad en la información en muchos de los campos importantes. Hasta ahora, no se ha realizado una gran cantidad de trabajos en el área de extracción automática de metadatos. Cada una de las herramientas para la extracción de metadatos tiene sus propios objetivos, arquitectura y técnicas de extracción. Aquí, analizamos algunos de estos sistemas, comparando cuatro de ellos: SAXEF, TWYS, Looking4LO, y MAGIC. Se presentaron brevemente dichos sistemas de extracción para luego compararlos de acuerdo a tres características: tipos de archivos, metadatos extraídos y recursos de procesamiento utilizados.

Hay otros sistemas que quedaron fuera del objeto de estudio dado que no se pudo encontrar información detallada acerca de ellos. Como por ejemplo OpenCalais[9] que es una herramienta que permite incorporar funcionalidad semántica dentro de blogs, sistemas manejadores de contenido, sitios Web o aplicaciones; y MetaGlance[10] que es un servicio Web utilizado para generar metadatos para páginas Web y documentos. Este último agrega en forma automática etiquetas a páginas Web, obtiene rápidamente la esencia de un documento, e incorpora los metadatos a un índice de búsqueda.

Consideramos que se debe incrementar el esfuerzo en la extracción automática de metadatos si se desea tener una correcta y completa información en los metadatos de los OA. En particular, la extracción de metadatos educacionales plantea nuevas dificultades. En esta dirección, existen una gran gama de líneas de investigación. Los sistemas deben ser capaces de tratar diferentes tipos de archivos, especialmente los no-estructurados, e

[9] http://www.opencalais.com/
[10] http://www.metaglance.com/

implementar técnicas híbridas que puedan ser aplicadas para la gran diversidad de metadatos.

## Agradecimientos

T. Pire agradece a LSIS por el apoyo brindado durante su estadía en el Laboratorio, donde comenzó el presente trabajo.

## Bibliografia

[1] Wiley, D., "Connecting Learning Objects to Instructional Design Theory: A definition, a metaphor, and a taxonomy", in D. A. Wiley (ed.) Instructional Use of Learning Objects. Editorial Association for Instructional Technology, 2002.

[2] Montaner, M., López, B.,de la Rosa, L. 2003. "A Taxonomy of Recommender Agents on the Internet". In Artificial Intelligence Review 19: 285-330, Kluwer

[3] Deco C., Bender C., Casali A., Motz R. "Design of a recommender educational system". In Proceedings 3ra. Conferencia Latinoamericana de Objetos de Aprendizaje LACLO 2008. México. pp 63-70. Octubre 2008.

[4] Bender, C., Motz, R., Deco, C., Saer, J. "Recuperación personalizada de e-cursos". In Proceedings IX Congreso Iberoamericano de Informática Educativa, RIBIE 2008. Caracas, Venezuela, marzo 2008.

[5] Casali, A., Gerling, V., Deco, C., Bender, C. "A Multiagent System for Personalized Recommendation of Learning Objects". Congreso Iberoamericano de Informática Educativa. IE 2010, Santiago de Chile, Diciembre 2010.

[6] Casali, A., Gerling, V., Deco, C., Bender, C., "Sistema Inteligente para la Recomendación de Objetos de Aprendizaje", LACLO 2010 pp.137-146. San Pablo, Brasil, Setiembre 2010.

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

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

[9] Alfano, M., Lenzitti, B., Visalli, N., "Text analysis module of a System for Automatic extraction of lEarning object Features (SAXEF)". In Proceedings III E-learning Conference, Coimbra, Portugal, Sept. 2006.

[10] Alfano, M., Lenzitti, B.,Visalli, N., "SAXEF: A System for Automatic eX-traction of learning object Features". Journal of e-Learning and Knowledge Society, vol. 3, (2), 83-92, 2007.

[11] Wai Yuen, T., "Automatic Extraction of Learning Object Metadata (LOM) from HTML Web Pages," Master of philosophy, City University of Hong Kong, May 2007.

[12] Motz, R., Badell, C., Barrosa, M., Sum, R., Díaz, G., Castro, M.: LooKIng4LO: Sistema Informático para la Extracción Automática de Objetos de Aprendizaje: Caso de Estudio. IEEE-RITA(2009) 223-229.

[13] Li, Y., Dorai, C., Farrell, R., "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, 2005.

[14] Rohlfing, K., Loehr, D., Duncan, D., Brown, A., Franklin, A., Kimbara, I., Milde, J., Parrill, F., Rose, T., Schmidt, T., Sloetjes, H., Thies, A., Wellingho, S., "Comparison of multimodal annotation tools workshop report", Gespraechsforschung, vol. 7, 99-123, 2006.