Prototipo informático para extracción de recursos digitales sobre internet

Computer prototype for digital resource extraction on internet

Paulo Alonso Gaona García
Ingeniero de Sistemas, magíster en Ciencias de la Información y las Comunicaciones, candidato a doctor en Ingeniería de la Información y del Conocimiento en la Universidad Alcalá de Henares. Madrid, España. Docente de la Universidad Distrital Francisco José de Caldas. Contacto: *pagaonag@udistrital.edu.co*

Salvador Sánchez Alonso
Ingeniero Informático, doctor en Informática. Docente Titular de la Universidad Alcalá de Henares. Madrid, España. Contacto: *salvador.sanchez@uah.es*

Eduardo Gaona García
Ingeniero Electrónico, magíster en Ciencias de la Información y las Comunicaciones. Docente de la Universidad Distrital Francisco José de Caldas. Bogotá, Colombia. Contacto: *egaona@udistrital.edu.co*

Fecha de recepción: 17 de abril de 2012
Fecha de aceptación: 16 de octubre de 2012
Clasificación del artículo: investigación
Financiamiento: Universidad Distrital Francisco José de Caldas - Universidad Alcalá de Henares

Palabras clave: AAT,Crawler Web, extracción de datos, Europeana, metadato, tesauro.

Key words: AAT, Crawler Web, data extraction, Europeana, metadata, thesaurus.

RESUMEN

El despliegue que ha tenido internet en los últimos años conlleva la creación de modelos de comunicación y de negocio que giran en torno a las actividades realizadas por una comunidad de tipo social, comercial, económico, académico o investigativo, permitiendo así interconectar millones de recursos informáticos sobre la Web mediante tecnologías todavía soportadas a través de HTML. El siguiente artículo presenta el modelo de desarrollo y operación de una herramienta de extracción de datos, para explorar recursos digitales de tipo Open Source a través de una fuente de recursos mediante una URL empleando como caso de estudio la Biblioteca Europeana. Se analizará su modelo de desarrollo, análisis de los resultados obtenidos y finalmente se presenta un apartado donde se enmarcan algunas tendencias de análisis y aplicaciones que se pueden abordar en el área de la educación.

ABSTRACT

The deployment that the Internet has had in recent years, involves the creationof business and communicationmodels and that revolve around the activities of a community of social type, commercial, economic, academic or research, allowing information resources interconnected million on the web using technologies yet supported by HTML. This paper presents the model development and operation of a extracting data tool for exploring Open Source digital resources through a resource via URL, «Case Study»: digital library Europeana. We will analyze its development model, analysis of results and finally presents a section where some are framed trends analysis and applications that can be addressed in the field of education.

* * *

1. INTRODUCCIÓN

Actualmente, se concibe la idea de internet como una red de comunicaciones que permite interconectar gran variedad de recursos y servicios, los cuales cada día se entrelazan a través de diversos tipos de tecnologías para su desarrollo y especificaciones para su despliegue. Existe una diversidad de recursos electrónicos sobre la Web que son fuente de información para su análisis en diversas áreas de conocimiento, bien sea para análisis estadístico, tendencias, comportamientos de mercado, consumo de productos, pronósticos, entre otros. Las herramientas de extracción Web son técnicas de análisis que permiten explorar diversas áreas de una página web y extraer aquellos elementos de interés para propósitos de estudio.

Este artículo pretende identificar procesos para llevar a cabo el desarrollo de un modelo de comunicaciones que permita explorar y extraer recursos de internet en un caso de estudio, la Biblioteca Europeana, mediante extracción de metadatos de cada uno de los recursos identificados a través de un conjunto de términos relacionados con el área de arte y arquitectura. Se muestran en la segunda sección, los trabajos relacionados en el campo de desarrollo de herramientas de extracción de datos y las diversas técnicas para su desarrollo. A continuación se presenta el modelo de trabajo planteado, identificando las fases de desarrollo de la herramienta. Finalmente, se presentan los análisis de los resultados obtenidos de la extracción y se incluye un apartado en el que se describen las tendencias de análisis y aplicaciones que se pueden implementar sobre el área de la educación.

2. INVESTIGACIONES RELACIONADAS

Crawler Web es una herramienta que permite analizar y extraer datos de una fuente de información sobre internet, para identificar sus características y elementos de desarrollo. A continuación se presenta un estudio de trabajos relacionados en el campo de la extracción de datos, las características y técnicas de su desarrollo y las fuentes de información usadas para la extracción.

2.1 Trabajos relacionados

El comienzo de los Crawler Web empezó en el año de 1993 en pleno desarrollo de internet. A partir de esta iniciativa se generaron desarrollos de herramientas más elaboradas y específicas para recorrer contenidos de información soportada en internet y la exploración de recursos, e indexación de los mismos mediante el desarrollo de motores de búsqueda y alternativas de clasificación de recursos sobre la Web [1]. Existe una

serie de iniciativas sobre el campo de la extracción de datos [2], [3], [4], propuestas orientadas hacia la arquitectura [5], a su estructura [6], escalabilidad [2], [7], [8], optimización [9], efectividad [10] y planteamiento de algoritmos de reconocimiento de patrones de extracción de alto nivel [11] implementados para mejorar el nivel de profundidad y análisis de estructuras de páginas Web. A continuación se relaciona una clasificación de estas herramientas [12] de acuerdo a su estrategia de uso.

- Crawlers en amplitud: orientados al desarrollo de motores de búsqueda y almacenamiento de archivos en internet [13], basados en la extracción de pequeños conjuntos de páginas web y los vínculos directos de cada una de ellas.
- Crawlers para actualización de páginas: verifican el índice de actualización de páginas web mediante varios rastreos realizados a través del Crawler. Esto permitió plantear estudios preliminares sobre técnicas de actualización de índices en la Web para indexar y mejorar estrategias de búsqueda en internet [1], [14] a través del historial de actualizaciones realizadas.
- Crawlers enfocados: utilizan motores de búsqueda más especializados para centralizarlos en temas específicos sobre ciertas páginas, esto permite enfocar la extracción sobre ciertos elementos como imágenes, audios, videos de varias URL, y bajo consumo de ancho de banda para el proceso. Se han desarrollado propuestas de extracción mediante esta estrategia basadas en análisis de estructura [15] y técnicas a través de Machine Learning [16], [17].
- Búsquedas aleatorias y muestreo: se han planteado técnicas [18], [19] orientadas a realizar búsquedas de manera aleatoria sobre márgenes de páginas web para identificar si se realizaron actualizaciones de contenido.

2.2 Estrategias de extracción de datos

La extracción de datos ha sido un campo de estudio que se ha desarrollado desde la inteligencia artificial, específicamente mediante la minería de datos. Desde sus comienzos ha permitido generar diversas técnicas mediante algoritmos computacionales que permiten analizar la información para generar conocimiento, abordando el campo de la extracción de conocimiento a partir de un conjunto de datos [20]. El descubrimiento de conocimiento sobre datos mediante herramientas de minería de datos, ha permitido combinar técnicas tradicionales de búsqueda, mediante numerosos recursos desarrollados en el área de la inteligencia artificial, matemáticas, estadística y la teoría de bases de datos [21] y, recientemente, sobre áreas de e-learning [22]. Estos comienzos dieron origen al desarrollo de técnicas orientadas a la Web mediante el uso de librerías especializadas para recorrer etiquetas y elementos representativos a partir de una URL, a estas herramientas se les conoce como Crawlers Web.

2.3 Herramientas de extracción Web

Existen proyectos orientados a la extracción mediante técnicas de Crawler Web que permiten extraer cierto tipo de información sobre páginas web a través de etiquetas HTML. Existe una variada lista de herramientas Open Source que permiten realizar esta tarea, este es el caso de Heritrix [23], el cual permite la extracción de ciertos tipos de recursos sobre una página Web; JSpider [24], desarrollado en lenguaje Java, que permite realizar ciertos recorridos de páginas web para ciertos elementos que se desean obtener de la URL, y otras herramientas relevantes como Arachnid [25], Web-Harvest [26], Crawler4j [27] y Ex-Crawler [28], las cuales presentan las mismas características. Sin embargo, son herramientas que, a pesar de ser

útiles para información específica, no tienen la versatilidad para adaptarse a grandes volúmenes de información, dado que manejan formatos de extracción planos, no tienen la posibilidad de manipular los datos para actividades de análisis y gestión en tiempo real, carecen de mecanismos para programar ciertas áreas de interés sobre una página, y la mayoría de herramientas extraen información de manera limitada.

2.4 Biblioteca Europeana

Europeana es un proyecto apoyado por la Unión Europea, cuyo propósito fundamental es ser la biblioteca digital abierta de patrimonio cultural de mayor cobertura sobre Europa. Cuenta con un amplio respaldo de proveedores y agregadores de contenidos en áreas relacionadas con el arte y patrimonio cultural a nivel europeo. Europeana está orientado a centralizar la mayor cantidad de recursos digitales alojados en repositorios externos para catalogarlos y facilitar el acceso a estos mediante la indexación de sus metadatos. Esta iniciativa ha permitido registrar una serie de catálogos gracias a la vinculación de más de 20 000 000 de recursos digitales al proyecto [29], de los cuales Europeana se encarga de definir lineamientos para el intercambio, normalización, almacenamiento, gestión y despliegue de los metadatos registrados.

2.5 Tesauro AAT

AAT (Art & Architecture Thesaurus) es un macrotesauro desarrollado por la Fundación Getty [30]. Su desarrollo está basado en las pautas establecidas por la organización internacional de normalización para la creación de tesauros monolingües ANSI/NISO z39.19 [31] y multilingües ISO 5964-1985 [32], en las cuales se destacan las características de indexación y recuperación de información. Su área de cobertura es el arte, la arquitectura y los materiales relacionados con el mundo cultural. Dentro del dominio de conocimiento, el tesauro AAT es uno de los más completos que se encuentran en el mercado [33], dado que maneja cerca de 131 000 términos definidos entre descriptores y citas bibliográficas almacenadas en una serie de registros. Su información se organiza a través de facetas y jerarquías.

3. METODOLOGÍA DE TRABAJO

Para el desarrollo de la herramienta fue necesario plantear una serie de actividades previas, por lo que se definieron 3 etapas básicas dentro del proceso, resumidas en: análisis, desarrollo y evaluación (figura 1). De esta manera, se trazaron las estrategias para el análisis semántico de la información a extraer y a nivel tecnológico para los requerimientos de la herramienta.

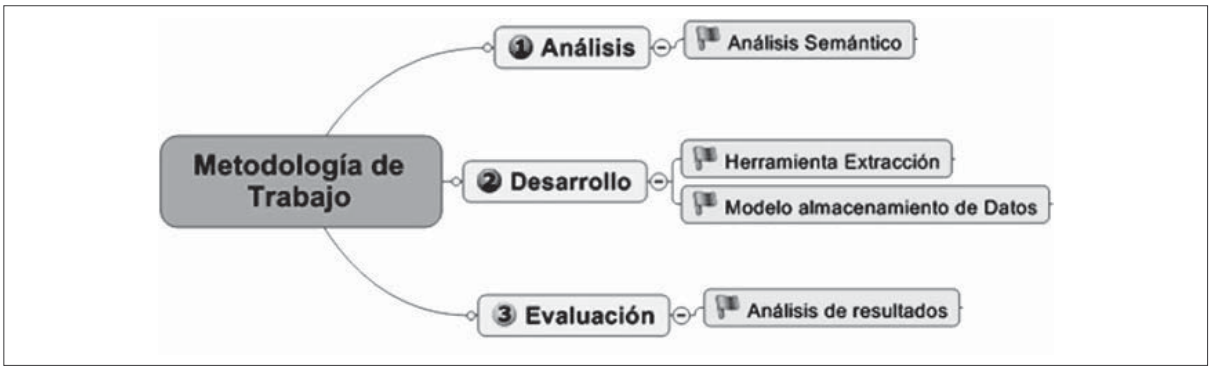

Figura 1. Modelo de trabajo
Fuente: elaboración propia.

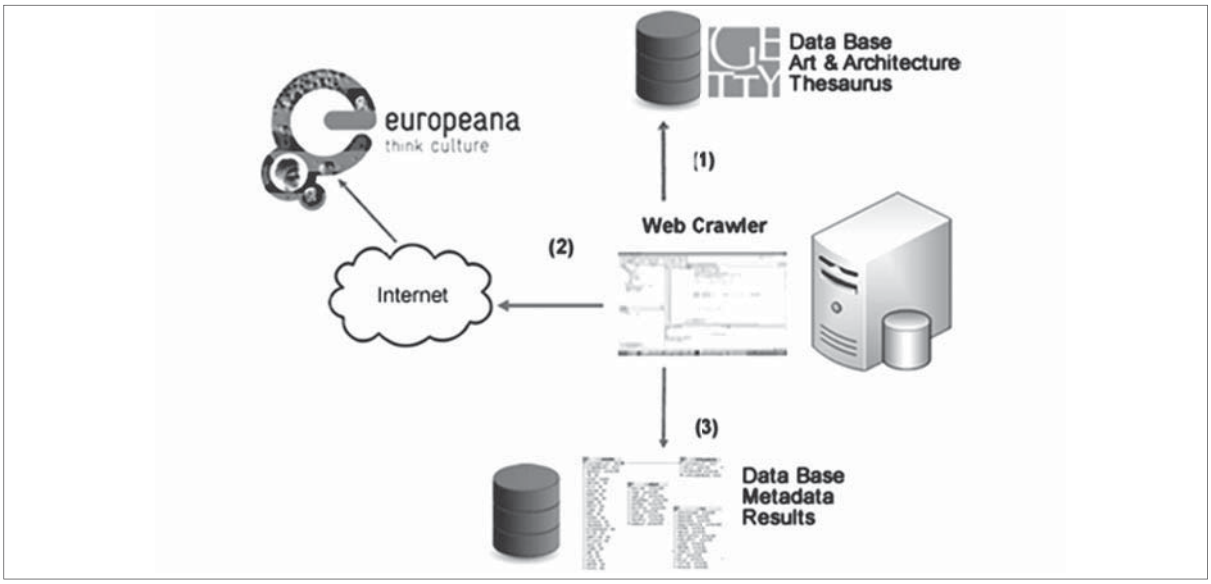

Figura 2. Modelo de extracción de datos
Fuente: elaboración propia.

3.1 Análisis semántico

La fase de análisis consiste en determinar la fuente de información de extracción de recursos; para ello, se tuvieron en cuenta los recursos relacionados a la Biblioteca Europeana, específicamente la extracción de sus metadatos definidos en su modelo de intercambio de datos ESE (Europeana Semantic Elements) [34].

3.2 Estrategia de extracción de datos

Se pretende realizar una exploración del número de recursos de Europeana relacionados a un conjunto de términos de conocimiento específico definidos a través del tesauro AAT (Art & Architecture Thesaurus). Por tanto, la herramienta de extracción deberá tomar un término con el propósito de lanzar consultas a la biblioteca de Europeana y, de acuerdo a los resultados arrojados, registrar el número de recursos digitales asociados al término y posteriormente realizar la extracción de cada uno de los recursos identificados a través de sus metadatos. A continuación se presenta el modelo de trabajo en la figura 2, la cual describe el proceso en los siguientes pasos:

1. Se realiza una consulta de los términos a explorar de acuerdo a una serie de términos almacenados en una base de datos del tesauro AAT.
2. Se identifica el número de recursos digitales de Europeana asociados a cada término.
3. Se extraen los recursos digitales identificados y se almacenan en una base de datos.

Básicamente, el Crawler Web realiza una consulta de una serie de términos del tesauro AAT alojados en una base de datos. Una vez identificados los términos a explorar, envía una petición de consulta del término sobre el buscador de Europeana.[1] A continuación, almacena el número de recursos digitales asociados al término y, posteriormente,

1 Disponible enhttp://www.europeana.eu/ [septiembre / 2012].

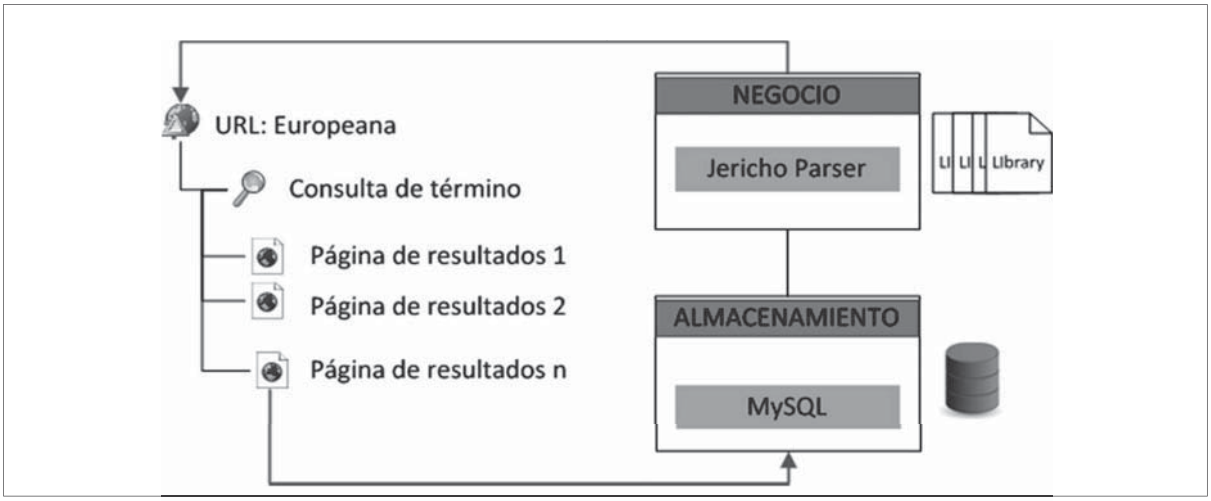

Figura 3. Arquitectura del Web Crawler
Fuente: elaboración propia.

el Crawler recorre cada resultado encontrado de cada paginación realizada por Europeana y realiza solamente la extracción de los metadatos de cada consulta lanzada a partir de los términos para su posterior almacenamiento. El proceso continúa hasta llegar al último término enviado para consulta de recursos.

3.3 Arquitectura del Crawler

Se utilizó un entorno de desarrollo en Netbeans sobre Java que básicamente dispone de dos capas de trabajo, representadas a través de una capa de negocio y otra de almacenamiento de registros encontrados, como se presenta en la figura 3.

Dentro de la capa de negocio se definieron estrategias de parseo de datos. Para su desarrollo se utilizaron estrategias de parseo de datos para limpiar la información extraída a través de la librería Jericho. Esta librería contiene métodos que permiten recorrer la página para identificar etiquetas de consulta, obtener resultados, almacenarlos en memoria, entre otros. Para este caso, se identificaron métodos para capturar el número de resultados del buscador de Europeana, recorrer las paginaciones realizadas por el motor de búsqueda de Europeana y características de exploración de metadatos de cada recurso digital identificado para su posterior extracción y almacenamiento.

3.3.1 Método para obtener el número de recursos

Este método permite lanzar una consulta sobre la página e identificar el número de resultados de acuerdo al término lanzado mediante el método getSource (tabla 1), y los almacena en una lista.

Tabla 1. Método getsource

```
public intgetNumberOfResults(String term)
{
int results = 0;
String resultsString;
try {
Source source = getSource("http://www.europeana.eu/portal/brief-doc.html?query=" + term + "&view=list");
List<Element> links = source.getAllElements("li");
for (Element link : links) {
String resultsOfQuery = "<li class=\"page-nr\">\nResults";
}
returnresults;
```

Fuente: elaboración propia.

4. MODELO DE ALMACENAMIENTO

Para el modelo de almacenamiento de datos, se trabajó con un modelo relacional en MySQL. Se tuvo en cuenta el manejo de registros que permitieran almacenar el número de coincidencias de los resultados obtenidos y los metadatos asociados a los recursos encontrados en Europeana. En la figura 4, se identifica el modelo entidad relación de la base de datos.

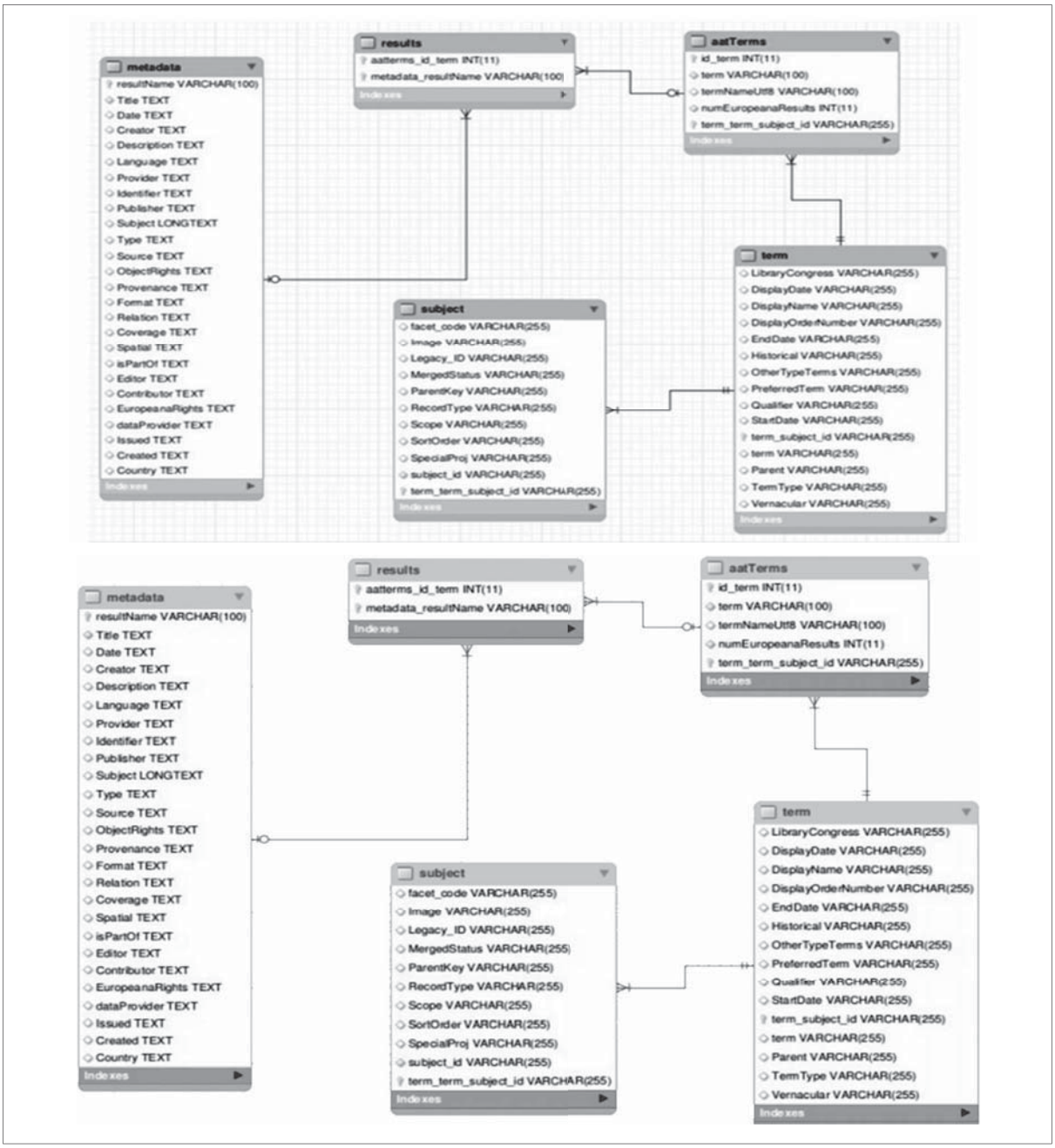

Figura 4. Modelo de almacenamiento
Fuente: elaboración propia.

A continuación se describen las tablas:

- Tabla AATterm: almacena el número de coincidencias del término que deseamos buscar.
- Tabla Results: se registran las URL identificadas por cada recurso en Europeana.
- Tabla Metadata: almacena los metadatos registrados de cada URL (recurso identificado), parámetros definidos en la especificación Europeana Semantic Elements (ESE).
- Tabla Term: se consultan los términos del tesauro.
- Tabla Subject: se relacionan las categorías de los términos del tesauro.

5. RESULTADOS OBTENIDOS

Se tomó como referencia un conjunto de 118 términos del tesauro AAT asociados a la faceta de Estilos y Periodos. Dentro del primer resultado obtenido, se identificaron cerca de 44 280 recursos digitales asociados a cada término en Europeana. Por cada recurso explorado se recorrió cada página para extraer sus metadatos. A continuación se realiza una descripción detallada de los recursos identificados.

5.1 Análisis de resultados extraídos

Mediante representación gráfica, se visualizan los cerca de 44 280 recursos digitales identificados en Europeana, los cuales se pueden identificar de manera proporcional en la figura 6, y discriminada por términos en la figura 7.

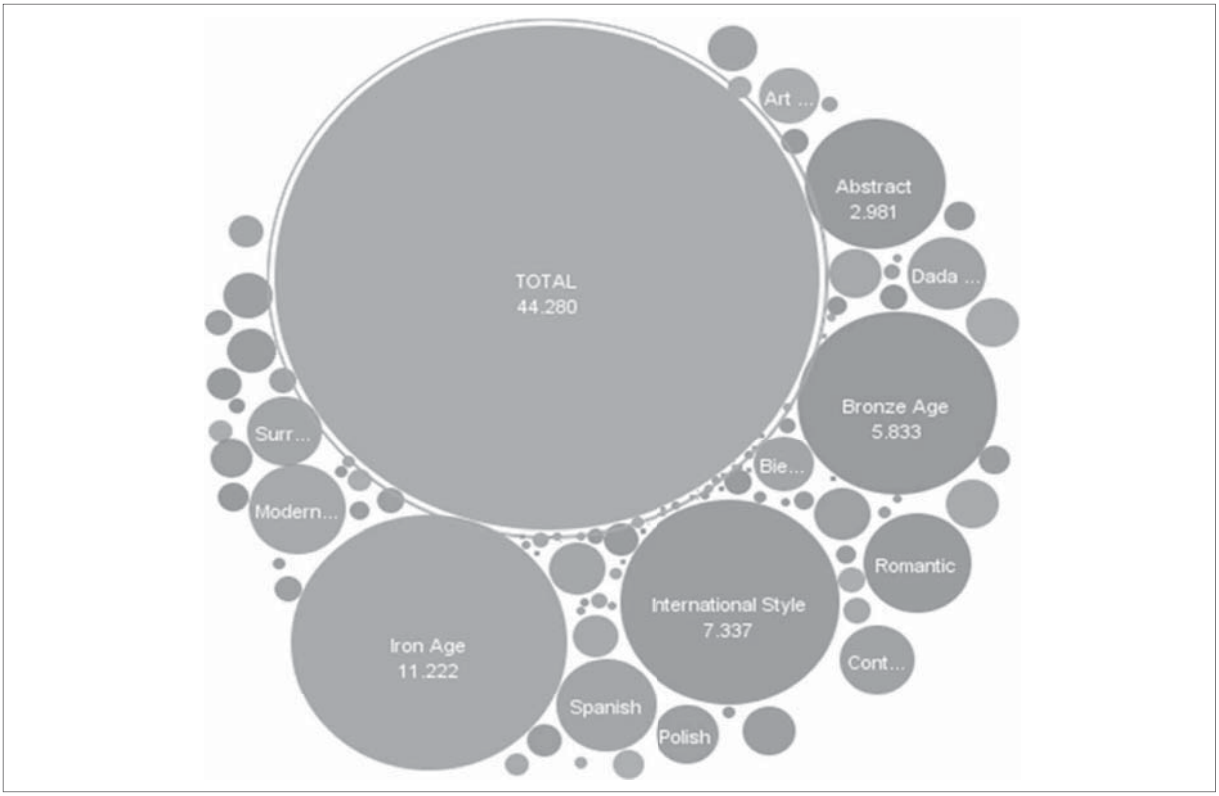

Figura 6. Total de recursos identificados en Europeana
Fuente: elaboración propia.

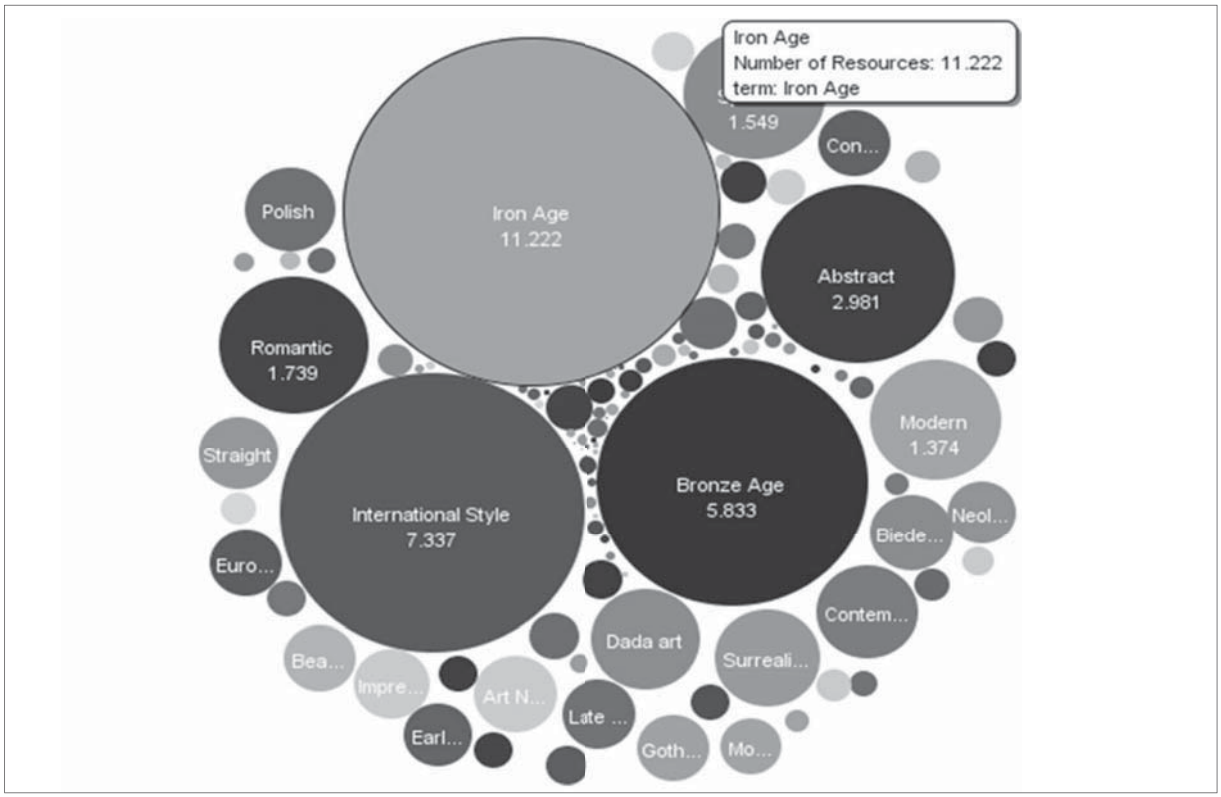

Figura 7. Número de recursos digitales por término
Fuente: elaboración propia.

En la figura 7 se presenta la discriminación de recursos digitales por término, identificándose tres términos predominantes: IronAge con un total de 11 222 recursos digitales para un 24% del total extraído, International Styles con un total de 7337 recursos digitales (16%) y BronzeAge con un total de 5833 recursos digitales (12%) del total de recursos explorados.

En la figura 8 se presenta el porcentaje de recursos digitales encontrados, discriminado por derechos de autor. Este análisis se realizó a partir de los registros de los metadatos de cada recurso digital explorado, encontrando que un 66% de recursos digitales presentan características de uso de libre acceso y un 4% de estos recursos poseen copyright. Sin embargo, un 30% de estos recursos no contenían descripciones de su uso, o simplemente no se encontraba categorizada.

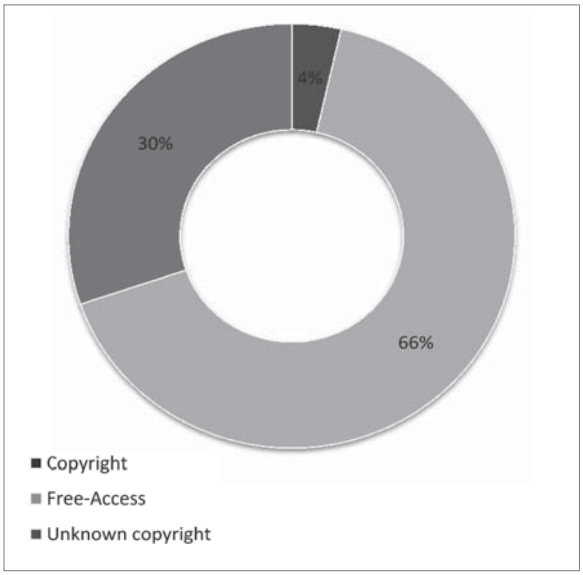

Figura 8. Porcentaje de recursos digitales por copyright
Fuente: elaboración propia.

6. ANÁLISIS DE RECURSOS EXTRAÍDOS VS. RECURSOS EXPLORADOS

Dentro del proceso de extracción de metadatos para identificar recursos digitales Open Source, se realizó un análisis del total de recursos explorados, es decir, los que se encontraron en Europeana vs. el total de recursos extraídos. Esto con el ánimo de analizar la fiabilidad de este tipo de herramientas, dado que hay factores que pueden generar un mal funcionamiento de este tipo de procesos, como el ancho de banda, el rendimiento de la máquina que realiza este proceso a nivel de procesamiento y de memoria y el número de recursos digitales que se desean extraer. En la figura 9 se presenta un acercamiento para analizar estas variables.

En la figura 9 en color gris oscuro se representa el número de recursos explorados vs. el número de recursos extraídos en color gris claro. Se puede identificar que de 44 280 recursos digitales explorados, se extrajeron 39 780, lo que le da una fiabilidad a la herramienta de un promedio de 89,84 % de efectividad en el proceso de extracción de recursos. A partir de estos resultados, se pueden identificar dos grandes diferencias relacionadas con los términos "*IronAge*" y "*Scottish*", resultados que pueden estar asociados a pequeños er-

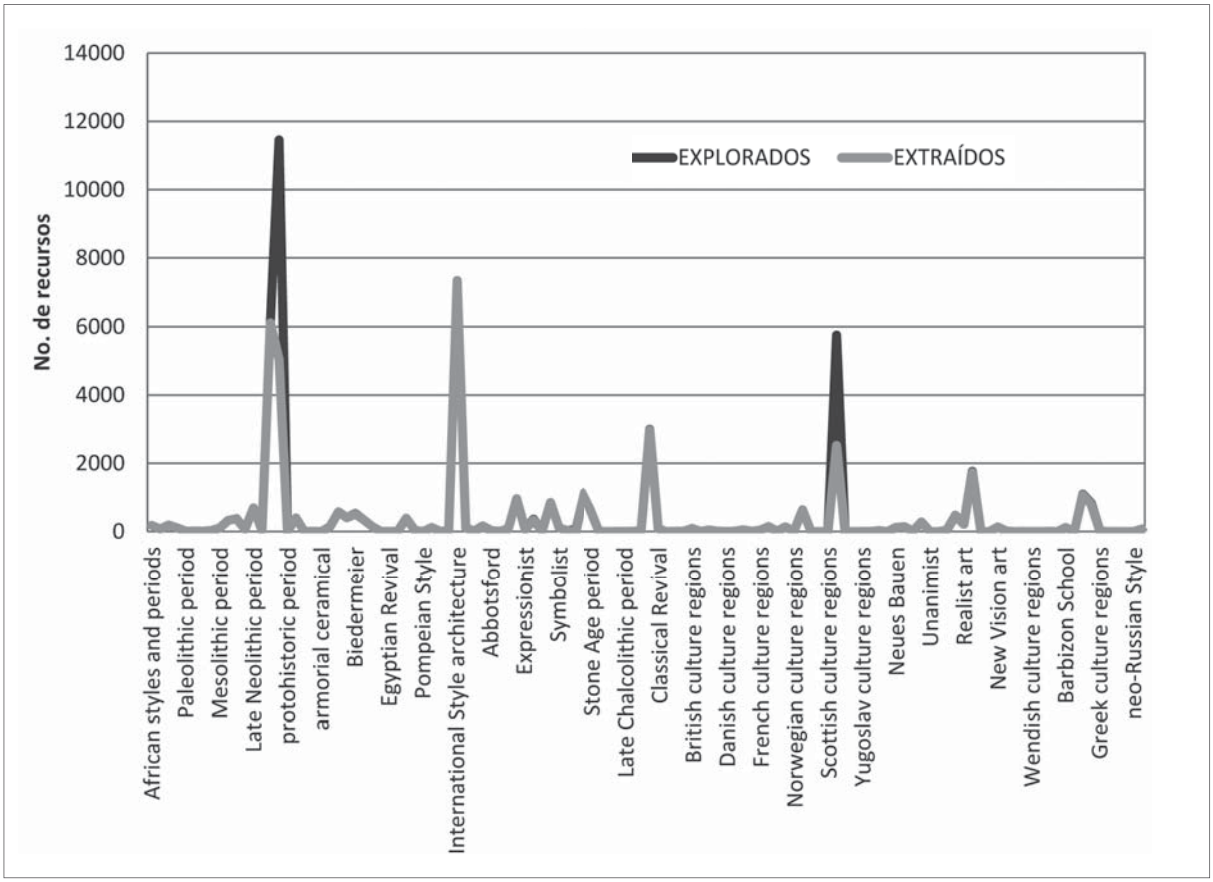

Figura 9. Número de recursos explorados vs. número de recursos extraídos
Fuente: elaboración propia.

rores de extracción relacionados con: i) recursos que presentaban características especiales con el tipo de vocabulario; ii) caracteres especiales del recurso digital; iii) mecanismos de paginación desarrollados por Europeana para clasificar y desplegar un gran volumen de recursos digitales identificados a partir de la consulta de un término.

7. APLICACIONES EN SECTOR ACADÉMICO

Dentro del campo de repositorios de objetos de aprendizaje, se podrían plantear estrategias de almacenamiento que permitan manejo e interoperabilidad de recursos digitales a través de la definición de modelos semánticos para el intercambio y vinculación de recursos digitales sobre diversos repositorios, permitiendo así la interoperabilidad de recursos digitales sobre la Web mediante estrategias de Linked Data [35].

A nivel de clasificación de recursos digitales, se podrían proponer modelos de búsqueda y categorización de contenidos de acuerdo a un área de conocimiento específico de la mano de herramientas como tesauros, que permiten definir una jerarquía de términos y conceptos asociados a un área de conocimiento específico.

El uso de este tipo de estrategias se puede implementar sobre repositorios de acceso libre, cuyas políticas de acceso permitan la exploración y uso de este tipo de mecanismos de extracción. Un caso específico sería a través del repositorio de recursos digitales que se encuentra desarrollando la Unión Europea bajo el proyecto ODS (Open Discovery Space)[36], una iniciativa que busca desarrollar e implementar un punto de acceso basado en la Web donde profesores de colegios de toda Europa pueden encontrar recursos educativos realizados por otros colegas o centros especializados con el fin de que, mediante el concepto de reutilización, puedan hacer uso de estos materiales en sus propias clases, adaptarlos a su contexto educativo y a las capacidades o estilos de aprendizaje de sus estudiantes para el desarrollo de objetos de aprendizaje.

8. CONCLUSIONES

El desarrollo de Web Crawlers para el caso de estudio abordado en este artículo, permite manejar buenos niveles de confianza de los recursos extraídos con un 86% de efectividad, lo cual para actividades de análisis de metadatos y desarrollo de objetos de aprendizaje permite contar con herramientas de primera mano para iniciar procesos de exploración y recuperación de información. A partir del número de recursos digitales identificados en el caso de estudio de Europeana, se contempla la posibilidad de definir estrategias para la clasificación de recursos con base en un área de conocimiento, mediante la ayuda de tesauros o modelos de representación de conocimiento definidos a través de ontologías.

Antes de realizar un proceso de extracción de recursos, es importante tomar el tiempo necesario para realizar un análisis previo sobre la URL a explorar, dado que este tipo de técnicas no las permiten algunas páginas Web debido a sus políticas de privacidad y uso de información.

Este tipo de mecanismos de extracción son adecuados para exploración de repositorios de acceso abierto que no dispongan de mecanismos para el intercambio de información a través de protocolos como Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) [37], tal como se mencionó en las aplicaciones de sector educativo a través del proyecto Open Discovery Space (ODS), iniciativa orientada a ofrecer una infraestructura para impulsar la adopción de recursos de aprendizaje en línea sobre colegios en toda Europa.

Los mecanismos de extracción de datos mediante Web Crawlers son métodos generales de extracción de recursos, pero no están exentos de dificultades. El proceso para llevar a cabo este tipo de actividades se enmarca en un análisis exhaustivo de la estructura semántica a partir de la URL a explorar, seguido de otros factores críticos como son el tiempo de extracción y las políticas de seguridad definidas por las páginas a explorar para permitir este tipo de actividades.

Una de las desventajas del desarrollo de herramientas de extracción de datos sobre internet tiene la característica de basarse en las estructuras semánticas de una página a partir de la definición de una URL. Si esta estructura cambia a través del tiempo, o se migra sobre otro tipo de lenguajes para el despliegue de los recursos sobre internet, las técnicas de Web Crawlers tienden a ser obsoletas, por lo que es necesario tener un amplio conocimiento a nivel de desarrollo sobre métodos y clases necesarias a partir de librerías especializadas para realizar el proceso de análisis y despliegue de la información de interés.

REFERENCIAS

[1] J. Cho and H. Garcia-Molina, *The evolution of the web and implications for an incremental crawler*, Stanford, 2000.

[2] A. Heydon and M. Najork, "Mercator: A scalable, extensible web crawler", *World Wide Web,* vol. 2, pp. 219-229, Springer 1999.

[3] R. D. Burke, "Salticus: guided crawling for personal digital libraries", in: *Proceedings of the first ACM/IEEE*-CS joint conference *on Digital Libraries, Roanoke, Virginia,* pp. 88–89, 2001.

[4] R. Baeza-Yates and C. Castillo, Balancing volume, quality and freshness in web crawling, In *Soft Computing Systems-Design, Management and Applications*, 2002.

[5] S. Chakrabarti*,* et ál., "Mining the Web›s link structure", *Computer,* vol. 32, pp. 60-67, 1999.

[6] P. Tadapak*,* et ál., "A machine learning based language specific web site crawler", in *Network-Based Information Systems (NBiS), 2010 13th International Conference on*, 2010, pp. 155-161, Takayama.

[7] P. Boldi*,* et ál., "Ubicrawler: A scalable fully distributed web crawler",*Software: Practice and Experience,* vol. 34, pp. 711-726, 2004.

[8] A. Tripathy and P. K. Patra, "A Web Mining Architectural Model of Distributed Crawler for Internet Searches Using PageRank Algorithm", in *Asia-Pacific Services Computing Conference, 2008. APSCC '08. IEEE*, 2008, pp. 513-518, Yilan: IEEE.

[9] J. Edwards*,* et ál., "An adaptive model for optimizing performance of an incremental web crawler", *Proceedings of the 10th international conference on World Wide Web*, 2001, pp. 106-113, New York: ACM.

[10] C. Castillo, "Effective web crawling", *ACM SIGIR* Forum, 2005, pp. 55-56, New York: ACM.

[11] P. Gupta and K. Johari, "Implementation of Web Crawler", in *Emerging Trends in Engineering and Technology (ICETET), 2009 2nd International Conference on*, 2009, pp. 838-843, Nagpur: IEEE.

[12] V. Shkapenyuk and T. Suel, *Design and implementation of a high-performance distributed web crawler*, 2002, pp. 357-368, San Jose, CA: IEEE.

[13] B. Kahle, «The internet archive Scientific American», Scientific American 1997.

[14] J.C. a. H. Garcia-Molina, Synchronizing a database to improve freshness In *Proc. of the ACM SIGMOD Int. Conf. on Management of Data,* pp. 117-128, 2000, New York: ACM.

[15] M.v.d.B.S. Chakrabarti, and B. Dom, "Focused crawling: A new approach to topic-specific web resource discovery", in *Proc. of the 8th Int. World Wide Web Conference (WWW8),* 1999, Elsevier.

[16] M. Diligenti, et ál., "Focused crawling using context graphs", 2000, pp. 527-534, Cairo, Egypt.

[17] J. R. a. A. McCallum, "Using reinforcement learning to spider the web efficiently.", In *Proc. of the Int. Conf. on Machine Learning (ICML),* 1999.

[18] A.H.M.R. Henzinger, M. Mitzenmacher, and M. Najork, "On near-uniform URL sampling", In *Proc. of the 9th Int. World Wide Web Conference,* 2000, Computer Network, Elsevier.

[19] A.H.M.R. Henzinger, M. Mitzenmacher, and M. Najork, "Measuring index quality using random walks on the web", In *Proc. of the 8th Int. World Wide Web Conference (WWW8),* pp. 213-225, 1999.

[20] U. Feyyad, "Data mining and knowledge discovery: Making sense out of data", *IEEE expert,* vol. 11, pp. 20-25, 1996.

[21] M. S. Chen, et ál., "Data mining: an overview from a database perspective", *Knowledge and data Engineering, IEEE Transactions on,* vol. 8, pp. 866-883, 1996.

[22] F. Castro, et ál., *Applying Data Mining Techniques to e-Learning Problems Evolution of Teaching and Learning Paradigms in Intelligent Environment*, vol. 62, L. Jain, et ál., Eds., ed: Springer Berlin Heidelberg, 2007, pp. 183-221.

[23] Heritrix, *Welcome to IA Webteam JIRA*, [Online]. Available:https://webarchive.jira.com/

[24] JSpider, *JSpiderinformation*, [Online]. Available: http://j-spider.sourceforge.net/

[25] Arachnid, *Arachnid information*, [Online]. Available: http://arachnid.sourceforge.net/.

[26] Web-Harvest, [Online]. Available: http://web-harvest.sourceforge.net/ [Accessed: December 2012].

[27] Crawler4j, Crawler4j information, [Online], Available: http://code.google.com/p/crawler4j/

[28] Ex-Crawler, Ex-Crawler information [Online]. Available: http://ex-crawler.sourceforge.net/joomla/

[29] Europeana, Europeana digital Library [Online], Available: http://pro.europeana.eu

[30] AAT, Art & Architecture Thesaurus (AAT) [Online], Available: http://www.getty.edu/research/tools/vocabularies/aat/

[31] ANSI/NISO, ANSI/NISO Z39.19 *Guidelines for the Construction, Format, and Management of Monolingual Controlled Vocabularies*, 2005.

[32] ISO, *ISO 5964-1985. Guidelines for the establishment and development of multilingual thesauri*, 1985.

[33] G. Mochón Bezares and Á. Sorli Rojo, "Tesauros de Humanidades en internet", *Revista española de documentación científica,* vol. 31, pp. 437-452, 2008.

[34] R.a.S. Clyphan, *Europeana Semantic Element ESE v3.4.1* [Online], Available: http://pro.europeana.eu/web/guest/technical-requirements/

[35] T. Berners-Lee, "Linked Data-The Story So Far", *International Journal on Semantic Web and Information Systems,* vol. 5, pp. 1-22, 2009.

[36] A. Lazonder, "ODS White Paper On the Adoption and Use of Elearning Resources Across Europe", *ODS Consortium,* 2012.

[37] C. Lagoze, et ál., "Open Archives Initiative Object Reuse and Exchange (OAI-ORE)", Technical report, Open Archives Initiative, 2007.