Desarrollo de un recomendador de metadatos para un repositorio utilizando técnicas de extracción de conocimiento
Developing a metadata recommender for a repository using knowledge extraction techniques

Alejandro Chuc Arcia y Víctor Hugo Menéndez Domínguez

Facultad de Matemáticas de la Universidad Autónoma de Yucatán, Anillo Periférico Norte, Tablaje Cat. 13615, Colonia Chuburná Hidalgo Inn, Mérida Yucatán. México
alexchuc84@gmail.com, mdoming@correo.uady.mx

Fecha de recepción: 28 de diciembre de 2020
Fecha de aceptación: 26 de abril de 2021

Resumen. Los repositorios digitales de producción científica son espacios que conservan la información de las diversas publicaciones almacenadas donde los metadatos juegan un papel fundamental para el cumplimiento de dicho propósito. Al describir a los recursos en términos de su contenido, utilización, características técnicas, etc., permiten su catalogación y, por ende, facilitan su localización, su recuperación y su uso. Sin embargo, los metadatos suelen adolecer de integridad, completitud, exactitud y consistencia. En este trabajo se propone desarrollar un asistente que facilite la generación de metadatos para los recursos almacenados en el repositorio DSpace utilizando técnicas de extracción de conocimiento. Se espera que las técnicas de extracción de conocimiento puedan facilitar y mejorar la generación de metadatos para trabajos de titulación almacenados en el repositorio DSpace, beneficiando especialmente a los usuarios no especializados en esta área. Otro resultado será una API basada en el modelo arquitectónico que facilite la generación automática de metadatos para los documentos almacenados en DSpace.

Palabras Clave: Repositorios Digitales, Generación de Metadatos, DSpace, Extracción de conocimiento.

Summary. The digital repositories of scientific production are spaces that preserve the information of the various stored publications where metadata play a fundamental role for the fulfillment of said purpose. By describing resources in terms of their content, use, technical characteristics, etc., they allow their cataloging and, therefore, facilitate their location, retrieval and use. However, metadata often suffers from completeness, completeness, accuracy, and consistency. In this work it is proposed to develop a wizard that facilitates the generation of metadata for the resources stored in the DSpace repository using knowledge extraction techniques. It is expected that knowledge extraction techniques can facilitate and improve the generation of metadata for titling works stored in the DSpace repository, especially benefiting users not specialized in this area. Another result will be an API based on the architectural model that facilitates the automatic generation of metadata for documents stored in DSpace.

Keywords: Digital Repositories, Metadata Generation, DSpace, Knowledge Extraction.

1 Introducción

En los últimos años, el concepto de ciencia abierta ha dominado el ambiente académico y científico, en este caso nos referimos a este modelo como una nueva concepción en el que se dice que los artículos científicos serán de acceso público, colaborativos y hechos por y para la sociedad [1]. Un elemento muy relacionado con la ciencia abierta son los repositorios institucionales, que son bases de datos documentales en las que se pueden realizar las acciones de consulta, carga y descarga de documentos académicos de una universidad o centro de investigación, por lo que se han convertido en un espacio relevante para la difusión del conocimiento almacenado [2].

Sin embargo, un problema siempre presente en un repositorio digital es la calidad de la información almacenada, eso por los procesos asociados a la captura de los descriptores del documento, especialmente la completitud y corrección de sus metadatos. Un metadato se define como un mapa, un tipo de estructura con significado por el cual la complejidad de un recurso se muestra de forma más simple, podemos pensar que es un elemento que da información y describe otro elemento informativo [3]. Los metadatos son usados para describir de forma breve a un documento, ejemplos de metadatos son el título, el resumen, el autor, las palabras clave, entre otros.

La completitud y la corrección de los metadatos garantiza la correcta búsqueda y recuperación de los documentos almacenados en el repositorio, lo que asegura su difusión y reutilización. La completitud se refiere a que los metadatos describan a los recursos de la manera más plena, esto es, el llenado de los datos usados para

describirlo se hace de la manera más íntegra, es decir, se centra en determinar cuantitativamente la calidad de los metadatos [4], mientras que el concepto de corrección puede ser descrito como la capacidad para determinar si un metadato cumple con las normas que le hayamos puesto, o si no, se modifique para que encaje de la manera adecuada en el modelo del sistema que se esté realizando [5].

En este sentido se propone una herramienta que facilite el proceso de incorporar metadatos a los documentos almacenados en el repositorio digital DSpace, de tal forma que garantice el cumplimiento de la completitud y corrección de sus metadatos utilizando técnicas automatizadas, estas técnicas se basan en la extracción del conocimiento, el cual es un proceso general deductivo que identifica patrones válidos, novedosos, útiles y comprensibles a partir de grandes y complejos volúmenes de datos [6]. Se propone utilizar técnicas de minería de texto, que consiste en la extracción de patrones útiles o conocimiento de documentos de texto [7] y aprendizaje automático, que es en un proceso general inductivo que crea automáticamente un clasificador aprendiendo las características de las categorías de un conjunto de documentos [8], para la generación de metadatos a partir del texto extraído en los documentos y metadatos almacenados en DSpace.

DSpace es un sistema de código abierto que funciona como un repositorio para las investigaciones digitales y material educativo producido por los miembros de universidad u organización [9].

Como caso de estudio se presentaría el uso de la herramienta dentro del repositorio digital de producción científica de una universidad pública para simplificar el proceso de captura de metadatos relacionados con los de trabajos de titulación de programas de posgrado.

2 Hipótesis y objetivo

¿Es posible facilitar la generación de metadatos de documentos almacenados en el repositorio Dspace usando técnicas de extracción de conocimiento?

Se pretende probar que las técnicas de extracción de conocimiento pueden hacer más sencillo y confiable la generación de metadatos de forma asistida para trabajos de titulación almacenados en el repositorio DSpace, especialmente para usuarios no especializados en estas áreas.

2.1 Hipótesis

Las técnicas de extracción de conocimiento pueden facilitar la generación de metadatos de documentos almacenados en el repositorio Dspace.

2.2 Objetivo general

- Desarrollar un asistente que facilite la generación de metadatos para recursos almacenados en el repositorio DSpace utilizando técnicas de extracción de conocimiento.

2.3 Objetivos específicos

- Identificar las propuestas relacionadas con la generación automática de metadatos, en particular aquellas relacionadas con técnicas de extracción de conocimiento, minería de texto y aprendizaje automático.
- Definir un modelo arquitectónico para la generación automática de metadatos con técnicas de extracción de conocimiento para documentos almacenados en un repositorio.
- Establecer criterios de calidad que garanticen el cumplimiento de la completitud y corrección de los metadatos asociados a un documento.
- Implementar una biblioteca de funciones que facilite la implementación del modelo arquitectónico propuesto donde se puedan validar las diversas prestaciones.
- Implementar un asistente generador de metadatos para el repositorio DSpace que utilice las bibliotecas desarrolladas para la generación automática de metadatos.
- Definir un conjunto de pruebas para evaluar el asistente desde la perspectiva funcional.
- Valorar la efectividad y eficiencia de la propuesta.

- Analizar el conjunto de datos arrojados durante la experimentación y pruebas.

3 Marco teórico

Debido al aumento en la cantidad de investigaciones y artículos desarrollados por las instituciones, muchas de ellas han optado por el uso de un repositorio institucional con lo cual se busca preservar dicha producción científica. Sin embargo, el proceso de publicación de un recurso en el repositorio es complejo, más aún para el usuario novato. Esto se debe principalmente a que muchas actividades involucradas pueden resultar tediosas en su realización o requerir conocimientos especializados.

A. Ciencia abierta y repositorios

La ciencia abierta abarca una multitud de supuestos sobre el futuro de la creación y divulgación de conocimiento [10]. Nos dice que la apertura de la ciencia, la investigación y la innovación a través de las tecnologías de la información y la comunicación (TIC) hace que la ciencia sea más eficiente, transparente e interdisciplinaria, y permite un mayor impacto social e innovación usando un nuevo enfoque del proceso científico basado en el trabajo cooperativo [11]. En este caso nos referiremos a la ciencia abierta como la práctica cuyo objetivo es incrementar y facilitar el acceso a las investigaciones científicas, materiales e información resultado de estos procesos, que hayan sido financiados con recursos públicos, con el propósito de que los ciudadanos se beneficien de la difusión máxima del conocimiento científico, tecnológico y de innovación [12].

Una parte importante para el éxito de la ciencia abierta es debido a los repositorios que cumplen con criterios de calidad y ofrecen opciones adecuadas de diseminación de contenidos y generalización de resultados de la investigación [13]. Un repositorio institucional es una base de datos cuya función es capturar, almacenar, ordenar, preservar y redistribuir la documentación académica de la universidad en formato digital [2].

Para la organización SPARC (Scholarly Publishing and Academic Resources Coalition), los repositorios institucionales que pertenecen a una institución son de ámbito académico, son acumulativos y perpetuos, y abiertos e interactivos [14].

B. DSpace

Debido al incremento de investigaciones y producción de material digital por parte de instituciones educativas, al igual que material tradicional, ahora investigadores y maestros elaboran recursos más complejos tales como audio, video, datos de aplicaciones heredadas, software y otro. En este sentido, DSpace ha sido desarrollado como un sistema para abordar esta necesidad de preservación digital, ofreciendo la funcionalidad requerida para un repositorio institucional a largo plazo de una manera simple [15].

DSpace es un sistema para repositorios que conserva, almacena, indiza y redistribuye material de investigación en formatos digitales de una organización [16]. Es una propuesta muy utilizada por sus diversas funcionalidades relacionadas con la gestión documental. Además de que está basada en estándares y tiene una arquitectura abierta y modular, lo que facilita su adecuación a necesidades específicas.

C. Generación de metadatos

Los metadatos son la información que creamos, almacenamos y compartimos para describir cosas y que nos permite interactuar con estas cosas para obtener el conocimiento que necesitamos [17].

Los metadatos tienen un papel importante al momento de fomentar la interoperabilidad y la reutilización entre distintas aplicaciones y contextos de aprendizaje, ya que describen los recursos en términos de su contenido, utilización, características técnicas, etc., permitiendo desarrollar servicios de catalogación, facilitar su localización, recuperación, entre otros [18].

La generación de los metadatos se puede producir de forma manual, automática o semiautomática utilizando diferentes técnicas de extracción de conocimiento [19]. Generalmente, los metadatos de un documento se almacenan en estructuras basadas en XML y conformes a un estándar, algunos de los más importantes para este proyecto son:

a. Dublin Core

Es uno de los estándares más usados en todo el mundo, se concentra en la descripción de las propiedades intrínsecas del recurso tales como el contenido intelectual o forma física, tiene como objetivo ser fácil de crear y mantener, permitir un entendimiento común en la semántica de los metadatos, considerar un ámbito internacional para la mejor representación de información y extender el conjunto base de las necesidades mediante perfiles de

aplicación para representar mejor las necesidades [20]. En este estándar todos los elementos son opcionales y no tiene un orden de aparición.

Este estándar contiene quince elementos de metadatos divididos en tres grupos:

- Contenido: título, descripción, fuente, idioma, relación, cobertura.
- Propiedad intelectual: autor, editor, colaborador, derechos.
- Instanciación: fecha, tipo, formato, identificador.

b. OpenAIRE

Es una infraestructura técnica que recolecta resultados de investigaciones de proveedores de datos conectados y define pautas de interoperabilidad de manera que sea compatible con este estándar, teniendo como objetivo establecer una infraestructura de comunicación académica abierta. Pretende proveer servicios de ciencia abierta, brindando servicios de interoperabilidad que conectan la investigación y permiten a los investigadores, proveedores de contenido, financiadores y administradores de investigación adoptar fácilmente la ciencia abierta. A través de la construcción de estándares comunes globales para vincular la investigación es posible el descubrimiento, transparencia, reproductibilidad y aseguramiento de la calidad de la investigación [21].

OpenAIRE extiende los metadatos del estándar Dublin Core para identificar recursos, proyectos, publicaciones y conjuntos de datos. Por otro lado, para recuperar los metadatos del conjunto de datos, OpenAIRE utiliza el protocolo OAI-PMH [22].

Los metadatos básicos que considera OpenAIRE para describir un recurso son: Title, Creator, Subject, Description, Publisher, Contributor, Date, Type, Format, Identifier, Source, Language, Relation, Coverage, Rights, Audience. Siendo la gran mayoría equivalentes a Dublin Core, de un total de 23. En OpenAIRE, los metadatos pueden ser obligatorios, obligatorios cuando corresponda, recomendados y opcionales.

D. Extracción de conocimiento (aprendizaje automático, minería de texto)

Debido a la necesidad de extraer información útil de los metadatos y recursos digitales almacenados en el repositorio, surge el tema de la extracción de conocimiento, que busca desarrollar métodos y técnicas para obtener conocimiento a partir de grandes cantidades de datos, esto debido a la dificultad para transformar fácilmente esos volúmenes de datos, en otras formas que pueden ser más compactas, abstractas o más útiles [23]. El proceso para la extracción de conocimiento es iterativo e interactivo, primero se desarrolla un entendimiento del dominio de la aplicación y el conocimiento relevante e identifica la meta del proceso, luego se elige un objetivo del conjunto de datos, para dar paso a la limpieza y preprocesamiento de los datos, y así realizar la reducción y proyección de los datos buscando características útiles con relación al objetivo, para establecer el objetivo del proceso con un método particular de minería de datos, por último se realiza el análisis exploratorio y la selección de modelos e hipótesis.

Dentro del proceso para la extracción de conocimiento, se utilizaría el aprendizaje automático, que consiste en generalizar comportamientos a partir de la información dada, ya que este método se centra más en la complejidad computacional del problema [24], para ofrecer una propuesta óptima de generación de metadatos con base a los datos existentes. Al igual que el aprendizaje automático, se requerirá analizar grandes cantidades de texto y descubrir nueva información de los datos mediante diferentes técnicas de minería de texto. La minería de texto se refiere a la recuperación de información, análisis de texto, extracción de información, categorización, agrupación y visualización de los datos [25].

4 Metodología

Se seguirá la metodología denominada investigación-acción, la cual, en síntesis, asocia la investigación con la práctica. La investigación informa la práctica y la práctica se encarga de informar la investigación de modo cooperativo [26].

Las fases son:

- Fase 1. Definición del escenario de la problemática y análisis del estado de arte. Revisión sistemática de la generación automática de metadatos, en particular aquellas relacionadas con técnicas de extracción de conocimiento, minería de texto y aprendizaje automático.
- Fase 2. Proponer las herramientas y las métricas que ayudarán a la solución de la problemática planteada. Estudio y definición de las herramientas que en conjunto pueden resolver la problemática, así como la definición de las métricas y técnicas que se utilizarán para la generación automática de metadatos.

- Fase 3. Implementar un prototipo para solucionar la problemática. De acuerdo con las herramientas propuestas, realizar el prototipo que podrá solucionar la problemática estudiada.
- Fase 4. Realización de pruebas. Comprobar que el prototipo tenga un correcto funcionamiento y usabilidad esperada, así como corregir anomalías.
- Fase 5. Documentación y difusión. Esta fase abarca todo el proceso de la tesis, ya que se documentará cada avance obtenido del trabajo, para que, en conjunto, sea el producto final de la tesis. Además, por cada avance obtenido en cada una de las fases mencionadas anteriormente, se comparten los resultados obtenidos con la comunidad científica a través de publicaciones.

5 Beneficio/Impacto

Se pretende generar diversos beneficios, los cuales se espera que tengan un impacto positivo en el área de la generación de metadatos.

Las contribuciones se enlistan a continuación en dos grupos: aportaciones teóricas, en forma de modelos; y aportaciones prácticas, dadas mediante componentes software desarrollados para la implementación de las aportaciones teóricas.

Aportaciones teóricas:

- Un modelo arquitectónico que facilite la generación automática de metadatos con técnicas de extracción de conocimiento para recursos de un repositorio.
- Un conjunto de indicadores utilizados para medir la calidad de los metadatos OpenAIRE en términos de completitud y corrección de los metadatos asociados a un documento.

Aportaciones prácticas:

- Una API basada en el modelo arquitectónico propuesto, que facilitará la implementación de generación de metadatos para los documentos almacenados en DSpace.
- Un asistente generador de metadatos que utilice las bibliotecas desarrolladas.

6 Conclusiones

La ausencia de herramientas que faciliten la generación de metadatos para los documentos almacenados en un repositorio digital es una problemática real. Esta actividad comúnmente se desarrolla en forma manual, lo que implica demasiado tiempo y esfuerzo humano, además que muchas veces es realizada por usuarios con pocos conocimientos en los temas, lo que la vuelve más propicia a errores humanos. Es posible facilitar este proceso a través de una herramienta computacional usando técnicas de extracción de conocimiento.

Para realizar este trabajo de investigación es fundamental estar familiarizado con el concepto de repositorios digitales, los cuáles son elementos comunes en la comunidad científica y académica. Otros temas asociados a la propuesta planteada son los metadatos y sus estándares, específicamente Dublin Core y OpenAIRE. Desde la perspectiva computacional la extracción del conocimiento y las técnicas de aprendizaje automático se consideran fundamentales para implementar la herramienta.

Se tiene contemplado que los beneficiados de la propuesta sean todos los usuarios no especializados en el proceso de publicación de recursos en un repositorio digital, especialmente en la generación de los metadatos de documentos.

Referencias

[1] L. Anglada and E. Abadal, "¿Qué es la ciencia abierta?," *Anu. ThinkEPI*, 2018, doi: 10.3145/thinkepi.2018.43.

[2] M. R. Barton and M. M. Waters, "Cómo crear un repositorio institucional: Manual LEADIRS II," *MIT Libr.*, no. Cmi, p. 169, 2005.

[3] J. Pomerantz, *Metadata*. MIT Press, 2015.

[4] V. H. Menéndez-Domínguez, M.-E. Castellanos-Bolañosa, C. Vidal-Castrob, and A. S. N, "Un Modelo de Calidad de Objetos de Aprendizaje basado en la Semántica de sus Metadatos," *Conf. LACLO*, vol. 3,

no. 1, p. 9, 2012.
[5] X. Zhao, H. Ma, H. Zhang, Y. Tang, and G. Fu, “Metadata extraction and correction for large-scale traffic surveillance videos,” in *Proceedings - 2014 IEEE International Conference on Big Data, IEEE Big Data 2014*, 2014, pp. 412–420, doi: 10.1109/BigData.2014.7004258.
[6] O. Maimon and L. Rokach, “Data mining and knowledge discovery handbook,” *Choice Rev. Online*, vol. 48, no. 10, pp. 48-5729-48–5729, 2011, doi: 10.5860/choice.48-5729.
[7] A.-H. Tan and others, “Text mining: The state of the art and the challenges,” in *Proceedings of the PAKDD 1999 Workshop on Knowledge Disocovery from Advanced Databases*, 1999, vol. 8, pp. 65–70.
[8] F. Sebastiani, “Machine Learning in Automated Text Categorization,” *ACM Computing Surveys*, vol. 34, no. 1. Association for Computing Machinery (ACM), pp. 1–47, 2002, doi: 10.1145/505282.505283.
[9] M. Smith *et al.*, “DSpace: An open source dynamic digital repository,” *D-Lib Mag.*, vol. 9, no. 1, 2003, doi: 10.1045/january2003-smith.
[10] B. Fecher and S. Friesike, “Open Science: One Term, Five Schools of Thought,” in *Opening Science*, S. Bartling and S. Friesike, Eds. Cham: Springer International Publishing, 2014, pp. 17–47.
[11] European Commission, “Open science: Political considerations from the European Commission,” 2016. [Online]. Available: https://pure.mpg.de/rest/items/item_2250860_3/component/file_2251123/content.
[12] CONACYT, “Repositorio Nacional.” [Online]. Available: https://www.repositorionacionalcti.mx/.
[13] T. Ferreras Fernández, “Visibilidad e impacto de la literatura gris científica en repositorios institucionales de acceso abierto. Estudio de caso bibliométrico del repositorio Gredos de la Universidad de Salamanca,” Universidad de Salamanca, 2016.
[14] SPARC, “SPARC: Advancing open access, open data, open education,” 2007. [Online]. Available: http://sparcopen.org/. [Accessed: 27-Jun-2020].
[15] R. Tansley, M. Bass, and M. Smith, “Dspace as an open archival information system: Current status and future directions,” in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2003, vol. 2769, pp. 446–460, doi: 10.1007/978-3-540-45175-4_41.
[16] S. E. Jaroszczuk, “Construcción de repositorios institucionales open source con Software Greenstone,” p. 120, 2010.
[17] J. (NISO) Riley, “Understanding Metadata - What Is Metadata?,” *Washingt. DC, United States Natl. Inf. Stand. Organ. (http//www. niso. org/publications/press/UnderstandingMetadata. pdf)*, p. 45, 2017.
[18] M. A. Sicilia and M. D. Lytras, *Metadata and semantics*. 2009.
[19] H. S. Al-Khalifa and H. C. Davis, “Folksonomies Versus Automatic Keyword Extraction: an Empirical Study,” *Proc. IADIS Web Appl. Res.*, vol. 1, no. 2, pp. 132–143, 2006.
[20] DCMI, “Dublin core metadata element set,” 2020.
[21] OpenAIRE, “OpenAIRE,” 2015. [Online]. Available: https://www.openaire.eu/.
[22] J. Corrales Correyero, “Directrices OpenAIRE 1.1: Directrices para proveedores de contenido del espacio de información OpenAIRE,” vol. 3, pp. 1–11, 2010.
[23] U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth, “From data mining to knowledge discovery in databases,” *AI Mag.*, vol. 17, no. 3, pp. 37–53, 1996.
[24] G. Pajares and J. de la Cruz, “Aprendizaje automático,” *Aprendiz. automático*, p. 376, 2011.
[25] S. Dang and P. H. Ahmad, “Text Mining : Techniques and its Application,” *Int. J. Eng. Technol. Innov.*, vol. 1, no. 4, pp. 22–25, 2014.
[26] D. E. Avison, F. Lau, M. D. Myers, and P. A. Nielsen, “Action research,” *Commun. ACM*, vol. 42, no. 1, pp. 94–97, 1999, doi: 10.1145/291469.291479.