Harvesting academic metadata through the OAI-PMH protocol to measure the impact of scientific publications

Rodrigo Velasco Luna
Facultad de Telemática
Universidad de Colima
Colima, México
rvelasco2@ucol.mx

Luis Gerardo Galindo Luna
Facultad de Telemática
Universidad de Colima
Colima, México
lgalindo2@ucol.mx

Pedro Cesar Santana Mancilla
Facultad de Telemática
Universidad de Colima
Colima, México
psantana@ucol.mx

Raymundo Buenrostro Mariscal
Facultad de Telemática
Universidad de Colima
Colima, México
raymundo@ucol.mx

José Román Herrera Morales
Facultad de Telemática
Universidad de Colima
Colima, México
rherrera@ucol.mx

Guillermo A. De la Torre Gea
Facultad de Telemática
Universidad de Colima
Colima, México
tmp_gtorre@ucol.mx

***Resumen*— El presente artículo tiene como objetivo probar la eficacia del protocolo para cosechado de metadatos de la iniciativa de archivos abiertos OAI-PMH (del inglés Open Archives Initiative - Protocol for Metadata Harvesting), se busca el cosechado de metadatos para construir una herramienta de búsqueda centrada en recursos de información académica en el contexto de un organismo público federal en México, con la finalidad de conocer el impacto de las publicaciones generadas por los investigadores mexicanos. En particular, el artículo, ofrece una visión general del desarrollo de la herramienta cosechadora de metadatos para la extracción de conocimiento en repositorios científicos. Los resultados obtenidos muestran que el protocolo OAI-PMH tiene potencial para búsquedas efectivas de información académica en repositorios abiertos.**

Palabras clave*— *Metadatos, OAI-PMH, Repositorios científicos, Impacto de publicaciones.

***Abstract*— The aim of this paper is to evaluate the effectiveness of the Open Archives Initiative protocol for the harvesting of metadata (OAI-PMH), in order to build a search tool focused on resources with academic information. All of this, in the context of a public federal organism in Mexico, to assess the impact of the publications generated by Mexican researchers. In particular, the article, offers an overview of the development of the metadata harvesting tool for the extraction of knowledge in scientific repositories. The results show that the OAI protocol has the potential for effective searches of academic information in open repositories.**

Keywords*— *Metadata, OAI-PMH, Scientific repositories, Publications impact.

I. INTRODUCCIÓN

El Instituto Nacional de Estadística y Geografía (INEGI) pone a disposición de la sociedad la información estadística y geográfica que produce. De esta forma contribuye al desarrollo del país ya que permite la toma de decisiones informadas, tanto por los gobiernos como por las organizaciones privadas y la academia. Debido a la importancia de dicha información, esta es usada frecuentemente como referencia en trabajos académicos y científicos como: tesis, artículos, libros, capítulos de libro, informes y reportes institucionales. Conocer el uso extendido de esta información que se da a través de la producción científica mundial, publicada a través de distintos repositorios y bases de datos heterogéneas, es de interés central del INEGI; por ello, se desea contar con un indicador del impacto de su uso (INEGI, 2017).

El Consejo Nacional de Ciencia y Tecnología (CONACyT), ha puesto a disposición de la comunidad científica su Repositorio Nacional1, el cual es una plataforma digital que proporciona acceso abierto en texto completo a diversos recursos de información académica, científica y tecnológica que producen con fondos públicos las instituciones en México. Uno de los objetivos principales de este repositorio es que debe ser abierto e interoperable (CONACyT, 2017). Para lograrlo, cuenta con un metabuscador que obtiene los datos de los recursos de información existentes en los repositorios registrados, los indexa y almacena sus metadatos con la finalidad de agregar estos recursos para poder ser cosechados por quien desee utilizarlos por medio del protocolo OAI-PMH (por sus siglas en inglés *Open Archives Initiative Protocol for Metadata Harvesting*) utilizando el esquema Dublin Core.

El protocolo OAI-PMH fue diseñado para compartir y describir recursos de información académica a través del web (Awre, 2006; Shreeves *et al.*, 2003). Surgió de la necesidad de investigadores y bibliotecarios de obtener fichas bibliográficas de las publicaciones disponibles en repositorios digitales, de una manera ordenada, estructurada y entendible por el usuario (Young, 2002). Esto les permitiría a distintas instituciones compartir publicaciones relevantes entre sí, y expandir sus

This article was accepted for presentation by an IEEE conference. Presented at IEEE ROC&C 2018-2019 without publication.

respectivas bibliotecas. Contrario a otros enfoques para buscar información distribuida, como el protocolo Z39.50, OAI-PMH cosecha los metadatos de cada proveedor y los agrega de forma central, lo que facilita el acceso describiendo los elementos en colecciones distribuidas y heterogéneas (Lagoze & Van de Sompel, 2001). Por lo tanto, la búsqueda es realizada por un proveedor de búsqueda en lugar de buscar de forma individual en cada repositorio.

Debido a lo anterior, se propone el uso del Repositorio Nacional de CONACyT para, por medio de su meta- buscador, tener acceso a los metadatos de todos los repositorios institucionales registrados a través del protocolo OAI-MPH. El presente artículo ofrece una visión general del desarrollo de una herramienta de búsqueda para cosechar información relacionada a las publicaciones del INEGI e integrar una base de datos con metadatos de referencias bibliográficas en los que se cita el uso de información de documentos estadísticos generados por dicho instituto para poder analizar y calcular en trabajo futuro el impacto de dichas publicaciones y en este primer paso probar la eficiencia del protocolo OAI-PMH en la tarea de recuperación de información.

II. Metodología

Como primer paso se definió la investigación, en esta se analizó el problema, se plantearon los objetivos, la justificación y el alcance del proyecto.

Posteriormente se definió como el lenguaje de programación a C# para desarrollar la herramienta de cosecha de metadatos por medio de OAI-PMH. Se logró acceder a las API (por sus siglas en inglés *Application Programming Interface*) del repositorio nacional de CONACyT y obtener los registros de la búsqueda de la palabra clave "INEGI" y sus combinaciones. La herramienta en su etapa de calibración regresó 22,183 resultados, sin términos de búsqueda, con estos registros es posible filtrar por palabras para refinar los resultados y realizar la cosecha de los documentos relacionados al INEGI, de los cuales se encontraron 51 documentos.

El software continúa con el procesamiento de los archivos XML (por sus siglas en inglés e*Xtensible Markup Language*) proporcionados como respuesta de la llamada a la API. XML es un metalenguaje de marcado que consiste en un conjunto de reglas simples para proveer un método uniforme para estructurar datos (Nicandro-Farías *et al.,* 2009). Una vez este proceso concluya, se genera una carpeta con un listado de todos los documentos encontrados y un archivo de texto plano que contiene los metadatos completos de los resultados encontrados.

A. *El protocolo Open Archives Initiative Protocol for Metadata Harvesting*

OAI-PMH es un protocolo aplicado principalmente a repositorios y bibliotecas digitales abiertas, con el propósito de ofrecer una manera sencilla de obtener metadatos de éstos. Se basa en un sistema petición-respuesta, en el que las peticiones son URLs estructuradas y las repuestas son documentos XML estandarizados (Purushothama & Bhandi, 2006). Las peticiones deben ser formadas de la siguiente manera:

[URL base del repositorio OAI] + [verbo] + [parámetro]

Los verbos, son las instrucciones con la petición al repositorio. Los verbos compatibles con el protocolo OAI-PMH, en su segunda versión, son:

- *ListMetadataFormats*: Enlista todos los prefijos de metadatos con los que es compatible el repositorio.
- *ListRecords*: Enlista todos los registros del repositorio. Como parámetro obligatorio, se debe especificar el prefijo de metadatos. Como parámetro opcional, se debe especificar un rango de fechas, un ID de conjunto, o un *ResumptionToken* para continuar procesando el resto de los recursos de información disponibles.
- *ListSets*: Regresa todos los conjuntos de registros disponibles en el repositorio.

B. *Cosecha por medio de ListRecords*

Como primer paso para realizar una cosecha, el programa obtiene un listado de todos los prefijos de metadatos disponibles en el repositorio, realizando una petición con el verbo *ListMetadataFormats*. Esta petición obtiene un listado de todos los prefijos que el repositorio soporta, y realiza el siguiente proceso en cada uno de ellos.

A continuación se muestra un ejemplo de una petición de tipo *ListMetadataFormats*:

https://oai-pmh.repositorionacionalcti.mx/resource/oai-pmh?verb=ListMetadataFormats

Se utiliza el verbo *ListRecords* para obtener los datos de los documentos disponibles en el repositorio. Es el método más directo y eficaz para recuperar la información de un repositorio. El cosechador realiza una petición con este verbo para obtener un listado de metadatos de los registros, y en caso de ser necesario, también obtiene un *ResumptionToken*, que confirma que el listado obtenido está incompleto, y puede ser utilizado en futuras peticiones para obtener la siguiente parte del listado, repitiendo el proceso las veces que sean necesarias. Cuando el valor obtenido del *ResumptionToken* es nulo, significa que el listado está completo.

En las siguientes peticiones se puede apreciar el uso del verbo *ListRecords*, sin usar el *ResumptionToken* y con él, respectivamente:

https://oai-pmh.repositorionacionalcti.mx/resource/oai-pmh?verb=ListRecords&metadataPrefix=oai_dc

https://oai-pmh.repositorionacionalcti.mx/resource/oai-pmh?verb=ListRecords&resumptionToken=8a0a47e6-333e-41a0-a28d-f7b31997a210

Es importante notar que cada *ResumptionToken* es un código único y solo puede ser utilizado una vez, además, tiene un periodo de validez que puede ser desde pocos minutos hasta horas, dependiendo del repositorio en el que se realice la cosecha. Este proceso se lleva a cabo utilizando todos los prefijos de metadatos disponibles en el repositorio.

C. *Cosecha por medio de ListSets*

Una cosecha con el verbo *ListSets* es muy similar al método que utiliza el verbo *ListRecords*, aunque requiere más tiempo de procesamiento y es menos directo. Una petición *ListSets* obtiene como respuesta un listado de todos los conjuntos de registros disponibles en el repositorio, de igual manera, este listado puede estar incompleto y contener un *ResumptionToken* para obtener la siguiente parte de la lista.

El cosechador realiza una "mini-cosecha" en cada conjunto utilizando una petición ListRecords especial, que obtiene todos los registros en ese conjunto únicamente, como se puede apreciar en el siguiente ejemplo:

https://oai-pmh.repositorionacionalcti.mx/resource/oai-pmh?verb=ListRecords&metadataPrefix=oai_dc&set=com_1011_1

Este proceso se repite para cada conjunto en el listado, y de igual manera, se repite por cada prefijo de Este proceso se repite para cada conjunto en el listado, y de igual manera, se repite por cada prefijo de metadatos disponible en el repositorio.

La razón por la que se implementaron los dos distintos métodos de cosecha fue que se comprobó que suelen obtener resultados distintos, dependiendo del repositorio en el que se realice la búsqueda de metadatos. Sin embargo, ambos métodos ofrecen al usuario resultados estructurados de la misma manera.

D. *Filtrado de resultados*

La principal debilidad del protocolo OAI-PMH es que no cuenta con un método de búsqueda por registros (excepto por fechas o periodos), por lo que el filtrado de resultados obtenidos se realiza de manera interna en el cosechador. El software recibe un término de búsqueda especificado por el usuario, realiza la cosecha de registros, y enlista solo aquellos que contengan el término en algún campo de sus metadatos.

Por este motivo, se considera un filtro de resultados, y no un método de búsqueda como tal.

III. RESULTADOS

Se desarrolló un prototipo en forma de un software de escritorio tipo cosechador (Figura 1) que hace uso de todas las funcionalidades que ofrece el protocolo OAI y que genera una base de datos que puede ser utilizada para los fines de este proyecto, de acuerdo con la propuesta metodológica de (Lagoze & Van de Sompel, 2001).

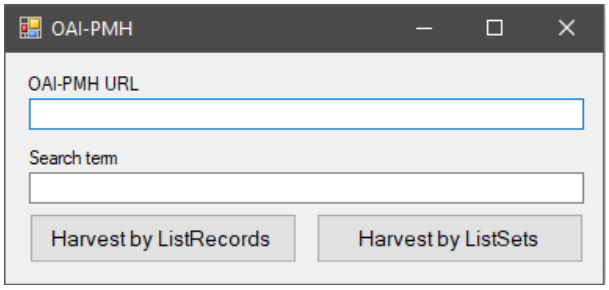

Figura 1. Diseño de la aplicación.

Como prueba, se realizó el *harvest* de tres repositorios a nivel nacional, los resultados se muestran en las Tabla 1 - 6.

- Repositorio de la Universidad de Colima.
- Repositorio Institucional del CIESAS.
- Repositorio Nacional CONACYT.

TABLA 1. REPOSITORIO DE LA UNIVERSIDAD DE COLIMA – LISTRECORDS.

Término de búsqueda	Registros encontrados	Registros eliminados	Registros totales	Páginas procesadas
Sin término	440	24	64	12
INEGI	0	No contabilizados	No contabilizados	12

TABLA 2. REPOSITORIO DE LA UNIVERSIDAD DE COLIMA – LISTSETS.

Término de búsqueda	Registros encontrados	Registros eliminados	Registros totales	Páginas procesadas
Sin término	880	48	928	12
INEGI	0	No contabilizados	No contabilizados	12

TABLA 3. REPOSITORIO INSTITUCIONAL DEL CIESAS– LISTRECORDS.

Término de búsqueda	Registros encontrados	Registros eliminados	Registros totales	Páginas procesadas
Sin término	1126	126	1152	2
INEGI	2	No contabilizados	No contabilizados	2

TABLA 4. REPOSITORIO INSTITUCIONAL DEL CIESAS– LISTSETS.

Término de búsqueda	Registros encontrados	Registros eliminados	Registros totales	Páginas procesadas
Sin término	20558	1625	22183	222
INEGI	55	No contabilizados	No contabilizados	222

TABLA 5. REPOSITORIO NACIONAL CONACYT – LISTRECORDS.

Término de búsqueda	Registros encontrados	Registros eliminados	Registros totales	Páginas procesadas
Sin término	20558	1625	22183	222
INEGI	55	No contabilizados	No contabilizados	222

TABLA 6. REPOSITORIO NACIONAL CONACYT – LISTSETS.

Término de búsqueda	Registros encontrados	Registros eliminados	Registros totales	Páginas procesadas
Sin término	21179	2458	23637	8
INEGI	96	No contabilizados	No contabilizados	222

Una vez que se realizaron las búsquedas, se contabilizaron los resultados con base al tipo de archivo (Carpenter, 2003), los cuales se muestran en la Figura 2, en donde la mayor parte fueron en formato PDF.

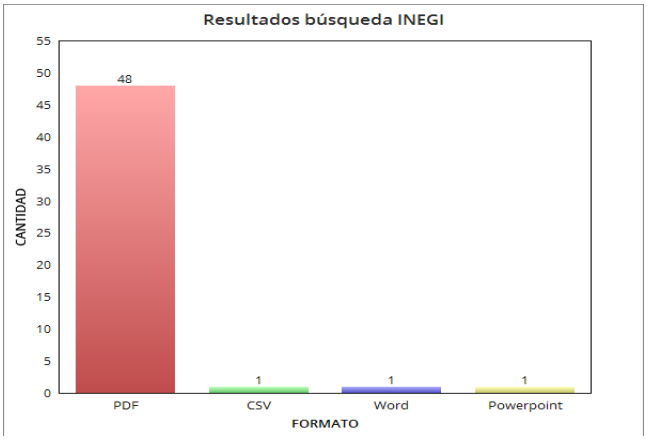

Figura 2. Tipos de archivo resultantes de la búsqueda.

IV. CONCLUSIONES

Con el constante crecimiento del volumen de información alojada en sistemas digitales, también ha incrementado la necesidad de organizar y cuantificar esta información. El protocolo OAI-PMH ha demostrado ser una solución eficiente a este problema, gracias a su facilidad de implementación y uso para bibliotecarios, investigadores, y estudiantes. El sistema de petición/respuesta basado en XML estándar también facilita su implementación para desarrolladores, pues obtener información de documentos XML es sencillo, y además, su transmisión es rápida, incluso en redes de baja velocidad. De las búsquedas realizadas se principalmente archivos en formato PDF, lo cual tiene una razón lógica debido a que los artículos científicos generalmente se almacenan en dicho formato. Se presentó una herramienta desarrollada para extraer información de repositorio de documentos académicos como tesis, artículos, libros, capítulos de libro, informes y reportes institucionales. Esta herramienta se desarrolló con la finalidad de identificar documentos relacionados con la información que genera INEGI, para medir el impacto de sus documentos estadísticos y geográficos. Para realizar la búsqueda y soportan el objetivo principal de que existe potencial para realizar búsquedas efectivas usando el protocolo OAI-PMH facilitando el análisis de la información obtenida para generar indicadores que permitan medir el impacto de su uso en publicaciones científicas y académicas de la información que genera el INEGI.

AGRADECIMIENTOS

Los autores agradecen al Fondo Sectorial CONACyT-INEGI por el financiamiento del proyecto 290379 de la convocatoria S0025-2016-2.

REFERENCIAS

[1] Ch. Awre. (2006). The technology of Open Access. En Neil Jacobs (Ed.), Open Access: Key Strategic,Technical and Economic Aspects (pp. 55-62). Oxford: Chandos Publishing. 2. CONACyT. (2017). Interoperabilidad con el Metabuscador del Repositorio Nacional. Recuperado de https://www.repositorionacionalcti.mx/docs/manualesInteroperabilidad/manual_de_Interoperabilidad_Re positorio_Nacional_ver.2.1.pdf

[2] CONACyT. (2017). Interoperabilidad con el Metabuscador del Repositorio Nacional. Recuperado de https://www.repositorionacionalcti.mx/docs/manualesInteroperabilidad/manual_de_Interoperabilidad_Re positorio_Nacional_ver.2.1.pdf 3. Gowda M Purushothama, & M K Bhandi. (2006). Metadata Harvesting and The Open Archives Initiative

[3] G. M. Purushothama, & M K Bhandi. "Metadata Harvesting and The Open Archives Initiative". En INFLIBNET's Convention Proceedings. INFLIBNET Centre, 2006.

[4] INEGI. (2017). Convocatoria FONDO SECTORIAL CONACYT-INEGI 2016-2.

[5] J.R. Young. "«Superarchives» Could Hold All Scholarly Output". The Chronicle of Higher Education, 2002, 48(10).

[6] C. Lagoze, & H. Van de Sompel. "The open archives initiative: building a low-barrier interoperability framework" 2001, (pp. 54-62). ACM Press. https://doi.org/10.1145/379437.379449

[7] M.E. Nicandro-Farias, P. Alcaraz, & L. Alcaraz. "Sistema Generador y Cosechador de Metadatos Dublin Core para Documentos de Tesis". En Octava Conferencia Iberoamericana en Sistemas, Cibernética e Informática: CISCI 2009. Orlando, FL, USA: International Institute of Informatics and Systemics.

[8] S.L. Shreeves, J.S. Kaczmarek, & T.W. Cole. "Harvesting cultural heritage metadata using the OAI Protocol". Library Hi Tech,2003, 21(2), 159-169. https://doi.org/10.1108/07378830310479802