

Nombre del proyecto: Generación de indicadores de productividad de cuerpos académicos y profesores.

Resumen:

Introducción

La investigación científica se ha vuelto más accesible gracias a los repositorios digitales, los cuales se encargan de almacenar, preservar y ser una herramienta de difusión de contenido. Los repositorios se pueden clasificar de acuerdo con su objetivo de la siguiente manera [1]: como de Recursos Digitales Abiertos (RDA), como de referencias a RDA, como iniciativas de OpenCourseware y como sistemas de gestión de aprendizaje. Un repositorio es una entidad de naturaleza digital o no digital, y puede ser utilizado, reutilizado o referenciado a lo largo del aprendizaje apoyado por la tecnología [2]. Existe la necesidad de describir y localizar los objetos de un repositorio, por lo que se hace uso de los metadatos. Éstos son datos descriptivos sobre los datos, es decir, proporcionan la información mínima necesaria para identificar un recurso [3]. Entre otras cualidades de los metadatos se encuentran [4]: el incremento al acceso a los RDA, disminución de tráfico en la red, expansión de la utilización de la información, control de versiones, aspectos legales y precisión de los procesos de búsqueda y recuperación.

En México, diversas instituciones cuentan con uno o varios repositorios digitales. A nivel nacional, el Repositorio Nacional es una plataforma digital la cual contiene en este momento 45 repositorios institucionales y proporciona acceso abierto en texto completo a una amplia gama de recursos de conocimiento académico, científico y tecnológico [5].

Otro caso de repositorio en México es la Red Mexicana de Repositorios Institucionales (REMEDI), compuesta por 39 instituciones de educación superior. Se enfoca en crear una red de interconexión de repositorios digitales de instituciones de educación superior del país para la integración, difusión, preservación y visualización de la producción científica, académica y documental generada en dichas instituciones, lo que le brinda la oportunidad de anexarse a redes o directorios de repositorios internacionales para el fomento del trabajo mutuo y apoyo al acceso y difusión de contenidos de acceso abierto [6].

Problemática

La calidad del trabajo de un investigador o de una comunidad académica o científica, está basada en la evaluación de los productos que genera y que comunmente son almacenados en repositorios digitales. Estos productos son valorados a partir de sus metadatos, su utilización, relevancia e impacto en el desarrollo de la ciencia. Dichas valoraciones pueden ser obtenidas mediante indicadores cuantitativos y estadísticos de

los documentos almacenados en los repositorios digitales. Asimismo, el grado de colaboración entre miembros de una comunidad científica puede determinarse a partir de la información registrada en el repositorio de productos generados por dicho grupo. Es así que se puede considerar que la calidad de la investigación está basada en la calidad de la información almacenada en el repositorio. Sin embargo, esta calidad puede verse comprometida por errores de captura, inconsistencia o ausencia de valores en los metadatos de los recursos.

En México, el Sistema Nacional de Investigadores (SNI) [7] y Programa para el Desarrollo Profesional Docente, para el Tipo Superior (PRODEP) [8], cuentan con criterios cualitativos propios para evaluar la calidad de los académicos de tiempo completo y de los cuerpos académicos que pueden verse beneficiados mediante de indicadores cuantitativos y objetivos, que pueden ser obtenidos de la información almacenada en los repositorios institucionales.

Antecedentes

Google Scholar es, a diferencia del buscador tradicional de Google, una herramienta de búsqueda de literatura del ámbito científico y académico. En él se pueden consultar títulos y autores. De hecho, Google Scholar permite a cualquier usuario visualizar los perfiles de los autores, donde se pueden observar las citas, el índice H, índice i10 (ambos indicadores se definen más adelante) y los coautores con los que ha trabajado [9].

Scopus es una base de datos de citas y resúmenes de literatura enfocada a la investigación. Al día 1 de junio de 2018 contaba con más de 25,000 títulos de más de 5,000 editoriales internacionales, 22,800 títulos seriales, 150,000 libros, 70 millones de artículos, 70,000 perfiles de instituciones y 16 millones de perfiles de autores. Su buscador permite realizar consultas por datos del autor (apellido, nombre y afiliación) [10] o bien, por ORCID, el cual es un identificador único para cada autor [11].

Microsoft Academic es un buscador académico basado en búsqueda semántica y no en palabras clave. Utiliza avances en aprendizaje automático, inferencia semántica y descubrimiento de conocimiento. Emplea el procesamiento del lenguaje natural para facilitar la comprensión y recordar el conocimiento transmitido en cada documento. Extrae el conocimiento de todas las publicaciones académicas descubiertas e indexadas por el buscador Bing. En los perfiles de autores de Microsoft Academic se encuentran los siguientes datos relacionados con su producción: número de artículos, número de citas, lista de coautores, afiliaciones de los coautores, temas de investigación del autor, así como nombre de las revistas y conferencias donde ha publicado. Al día 4 de mayo de 2018, Microsoft Academic contaba con 173,763,013 publicaciones, 228,528 campos de estudio, 4027 publicaciones de conferencias, 210,806,214 autores, 47,939 revistas y 18,719 afiliaciones [12].

En México existe el Repositorio Nacional, el cual se define como una plataforma digital encargada de proporcionar acceso abierto (es decir, sin requerimientos de suscripción, registro o pago) en texto completo a una amplia variedad de recursos de información de tipo académica, científica y tecnológica. Entre los materiales aceptados tanto para el Repositorio Nacional y los repositorios institucionales se encuentran tres tipos de Recursos de Información Académica, Científica, Tecnológica y de Innovación [13]:

- **Publicaciones científicas.** Universo de publicaciones provenientes de la investigación, es decir, artículos, libros, capítulos de libros, tesis de posgrado y disertaciones, memorias de congresos y reportes de investigación, entre otros.
- **Productos de desarrollo tecnológico e innovación.** Se refiere a patentes, desarrollos tecnológicos, innovaciones, transferencias tecnológicas, mejoras o diagnósticos del estado de la tecnología y mejoras o creaciones de prototipos, procesos o servicios que no se encuentren o son deducidos del estado de la tecnología.
- **Datos de las investigaciones.** Información recolectada y utilizada para fines de la investigación académica, científica, tecnológica y de innovación comúnmente aceptada por la comunidad científica como un elemento vital para la validación de resultados de las investigaciones.

Al día 1 de noviembre de 2017, el Repositorio Nacional estaba compuesto por 33 repositorios institucionales, 10,797 recursos de información y 282,355 consultas, mientras que para el 1 de junio de 2018 contaba con 45 repositorios institucionales, 18,153 recursos de información y 625,392 consultas, existiendo un incremento de 36.3%, 68.1% y 121.4% respectivamente en un lapso de 7 meses. El Repositorio Nacional permite realizar búsquedas simples y avanzadas. La segunda puede ser parametrizada o por área de conocimiento.

La búsqueda avanzada parametrizada puede contener:

- Palabras
- Título
- Autor
- Colaborador
- Materia
- Tipo
- Editor
- Nivel de acceso
- Repositorio
- Fecha inicial

La búsqueda por área de conocimiento se refiere a las siguientes:

- Biología y Química
- Ciencias Agropecuarias y Biotecnología
- Ciencias Físico- Matemáticas y Ciencias de la Tierra
- Ciencias Sociales
- Humanidades y Ciencias de la Conducta
- Ingeniería y Tecnología
- Medicina y Ciencias de la Salud

Al consultar los datos de un autor, el repositorio digital, únicamente proporciona los productos del autor, el tipo de producto (artículo, patente, tesis, etc.) y al área de conocimiento que pertenece cada producto [5]. No presenta indicadores de citas, como Google Scholar.

Los Repositorios Institucionales de México tienen como finalidad el acopio, preservación, gestión y acceso electrónico de información y contenidos de calidad, incluyendo aquellos de interés social y cultural que produzcan las instituciones, centros de investigación y demás integrantes del Sistema Nacional de Ciencia, Tecnología e Innovación de México, que hayan sido financiados de forma total o parcial con recursos públicos o bien, hayan utilizado infraestructura pública [13].

Muchos de estos repositorios se agrupan en REMERI. No obstante, las búsquedas realizadas a través de esta red de repositorios, únicamente visualizan productos de acuerdo con su tipo (tesis, artículo, etcétera), título, autor, adscripción y dirección en Internet (o URL). No dan a conocer indicadores del autor, ni algún otro dato relevante. Los repositorios contienen productos, los cuales se describen con metadatos. Un metadato es un dato con una estructura que está por encima de la información, es decir, son los datos que describen a los datos a analizar [14]. Por otra parte, los metadatos se estructuran de tal manera que facilitan la administración, el descubrimiento y la recuperación de los recursos [15] y, es debido a la necesidad de garantizar la calidad de los metadatos, que se han propuesto mejoras, sobre todo en el cumplimiento de la funcionalidad [16]. El uso de los metadatos debe considerarse primordial para la obtención de nuevo conocimiento, ya que de lo contrario esto ocasionaría búsquedas ineficientes de información, y por consecuencia, falta de aprovechamiento del conocimiento [17]. Es por medio de los repositorios digitales y los metadatos que se desean calcular los indicadores cuantitativos y estadísticos de la producción científica. Algunos de ellos se mencionan a continuación:

- **Factor de Impacto de la Revista (FIR).** Consiste en la proporción de citas en el año actual referentes a los artículos publicados en la revista en los últimos dos años (A) dividido entre el número de artículos publicados en los mismos dos años (B) [18]. La fórmula es la siguiente:

$$FIR = \frac{A}{B}$$

- **Citas.** Para calcular el rendimiento del trabajo de investigación de una revista científica o comunidad científica, se utilizan fórmulas, donde en todos los casos, N es el número total de publicaciones de la comunidad científica o revista, FIR_i representa el FIR que contiene la i -ésima publicación, mientras que Cn_i corresponde al número total de citas de la i -ésima publicación. Se toman los valores de los últimos dos años [18]. $SFIR$ es la suma de todos los FIR de todas las publicaciones. SC es la suma de todas las citas que se han hecho a las publicaciones de la comunidad científica o revista. IdI es el Impacto de la Investigación [18].

La fórmula de $SFIR$ es:

$$SFIR = \sum_{i=1}^N FIR_i$$

SC se obtiene a través de la siguiente fórmula:

$$SC = \sum_{i=1}^N Cn_i$$

IdI se calcula a través de la fórmula que se presenta a continuación:

$$IdI = \sum_{i=1}^N FIR_i (1 + Cn_i)$$

- **Índice H.** Aplica para la evaluación individual de un investigador, más que a una comunidad científica. N_p es el total de trabajos de un investigador. Un investigador tiene como índice el valor h si h de sus N_p trabajos cuentan con, por los menos, h citas. Los trabajos $N_p - h$ cuentan con h citas como máximo [19]. Por ejemplo, un investigador tendrá un *índice H* de 8, cuando 8 de sus artículos hayan sido citados por lo menos 8 veces cada uno.
- **Índice i10.** Indicador creado y utilizado por Google. Representa la cantidad de publicaciones con al menos 10 citas [20].

Propuesta

En esta propuesta se plantea la creación de una herramienta tecnológica que determine la calidad del trabajo de un académico utilizando los metadatos de su producción académica y científica que son almacenados de repositorios digitales, por medio de indicadores cuantitativos y estadísticos para así obtener datos relevantes, así como su nivel de producción interna (colaboraciones con académicos de la misma institución) y externa (colaboraciones con académicos de otras instituciones).

Objetivo general:

Proponer una solución tecnológica que determine la calidad del trabajo de un académico, a través de la evaluación de los metadatos de su producción académica y científica, los cuales se podrán consultar en uno o más repositorios digitales.

Objetivos específicos:

- Realizar una revisión de las herramientas existentes que den a conocer los indicadores cuantitativos y estadísticos de la producción académica y científica de un académico.
- Establecer las herramientas y técnicas que realicen un análisis profundo de los datos.
- Aplicar un algoritmo de Aprendizaje Automático que clasifique las variantes del nombre completo de un autor y corrija sus inconsistencias.
- Implementar una solución tecnológica que calcule la calidad del trabajo de los académicos, por medio de indicadores cuantitativos y estadísticos.
- Definir métricas que midan la calidad de la solución tecnológica.
- Analizar experimentalmente el funcionamiento de la solución tecnológica con el Repositorio Digital Institucional de la Universidad Autónoma de Yucatán.

Metodología de la investigación:

Las fases consideradas en la metodología de la investigación son:

- **Fase 1. Definir el escenario de la problemática y análisis del estado de arte.** Se refiere a mejorar la recolección de la información acerca de la producción académica y científica de un profesor almacenada en repositorios digitales. Esto origina la necesidad de definir una herramienta tecnológica que permita por medio de los metadatos asociados a los recursos generados por un académico, medir la producción científica por medio de indicadores cuantitativos y estadísticos. Se realiza el análisis del estado del arte, el cual consiste en hacer una revisión sistemática de las temáticas afines a los repositorios digitales, los metadatos (incluyendo su calidad y estándares), la interoperabilidad entre repositorios, los problemas existentes en las

comunidades científicas en relación con los metadatos, así como los indicadores cuantitativos y estadísticos para medir la producción científica, además del análisis de las evaluaciones a los investigadores por parte de las instancias, las cuales requieren de los datos de producción científica.

- **Fase 2. Proponer las herramientas y las métricas que ayudarán a la solución de la problemática planteada.** Consiste en definir el modelado de la solución tecnológica por medio de herramientas especializadas, establecer las herramientas de software que permiten realizar de manera eficaz y persistente la solución planteada, así como la definición de las métricas para la evaluación del desempeño de la misma.
- **Fase 3. Implementar un prototipo para solucionar la problemática.** A partir del modelado, desarrollar una herramienta de software que permita obtener los metadatos de toda la producción científica de cada profesor en cuestión para su posterior análisis y obtención de los indicadores de producción, así como diseñar el front- end que presenta los resultados de forma comprensible al usuario.
- **Fase 4. Realización de pruebas.** Por medio de las métricas establecidas en la fase 2, se realizan estudios que determinan el nivel de desempeño de la herramienta. Se evalúa la efectividad y la eficiencia del uso de la herramienta en escenarios reales.
- **Fase 5. Documentación y difusión.** Se comparten los resultados obtenidos en las fases anteriores a la comunidad investigadora por medio de la presentación de trabajos en revistas científicas y en foros nacionales e internacionales.

Cronograma:

Inicia en enero de 2018, termina en junio de 2019. Los números relacionados a cada año corresponden al número de mes.

	2018												2019					
Fase	1	2	3	4	5	6	7	8	9	10	11	12	1	2	3	4	5	6
1	■	■	■	■														
2				■	■	■	■											
3							■	■	■	■	■	■						
4													■	■	■			
5	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■

Referencias:

- [1] X. Ochoa y E. Duval, "Quantitative analysis of learning object repositories", *IEEE Trans. Learn. Technol.*, vol. 2, no. 3, pp. 226–238, 2009.
- [2] A. Ip, I. Morrison y M. Currie, "What is a learning object, technically", *Proc. WebNet*, pp. 23–27, 2001.
- [3] J. A. Senso y A. de la Rosa Piñero, "El concepto de metadato: algo más que descripción de recursos electrónicos", *Ciência da Informação*, vol. 32, no. 2, pp. 95–106, 2003.
- [4] M. Agudelo, "Los metadatos", *Gestión Contenidos Educ. Virtual Calid*. Ministerio de Educación Nacional de Colombia, Antioquía, Colombia, p. 5, 2009.
- [5] Consejo Nacional de Ciencia y Tecnología, "Repositorio Nacional/Sección Informativa", 2017. [En línea]. Disponible en: <https://www.repositorionacionalcti.mx/>. [Acceso: 30-Nov-2017].
- [6] Red Mexicana de Repositorios Institucionales, "Portal REMERI - Información", 2017. [En línea]. Disponible en: <http://www.remeri.org.mx/portal/acerca.html>. [Acceso: 25-Oct-2017].
- [7] Consejo Nacional de Ciencia y Tecnología, "Sistema Nacional de Investigadores", 2017. [En línea]. Disponible en: <https://www.conacyt.gob.mx/index.php/el-conacyt/sistema-nacional-de-investigadores>. [Acceso: 20-Nov-2017].
- [8] Dirección General de Educación Superior Universitaria, "Programa para el Desarrollo Profesional Docente, para el Tipo Superior (PRODEP)", 2017. [En línea]. Disponible en: <http://www.dgesu.ses.sep.gob.mx/PRODEP.htm>. [Acceso: 25-Oct-2017].
- [9] Google, "Google Académico", 2017. [En línea]. Disponible en: <https://scholar.google.com.mx/>. [Accesed: 12-Dic-2017].
- [10] Elsevier B.V., "Scopus", 2018. [En línea]. Disponible en: <https://www.scopus.com/freelookup/form/author.uri>. [Acceso: 13-Dec-2017].
- [11] Open Researcher and Contributor ID, "ORCID. Nuestra Misión", 2017. [En línea]. Disponible en: <https://orcid.org/about/what-is-orcid/mission>. [Acceso: 21-Nov-2017].
- [12] Microsoft, "Microsoft Academic. Frequently Asked Questions", 2018. [En línea]. Disponible en: <https://academic.microsoft.com/#/faq>. [Acceso: 04-May-2018].
- [13] Consejo Nacional de Ciencia y Tecnología, "Lineamientos Técnicos para el Repositorio Nacional y los Repositorios Institucionales", 2014. [En línea]. Disponible en: <http://www.siicyt.gob.mx/index.php/normatividad/2-conacyt/4-conacyt/1499-lineamientos-tecnicos-para-el-repostitorio-nacional-y-los-repositorios-institucionales/file>. [Acceso: 17-Ene-2018].
- [14] M. J. Lamarca Lapuente, "Hipertexto: El nuevo concepto de documento en la cultura de la imagen", Lamarca Lapuente, 2013.
- [15] H. S. Al-Khalifa y H. C. Davis, "The evolution of metadata from standards to semantics in E-learning applications", en *Proceedings of the seventeenth conference on Hypertext and hypermedia - HYPERTEXT '06*, 2006, p. 69.
- [16] C. Vidal-Castro, A. A. Segura-Navarrete, V. H. Menéndez-Domínguez y C. Martínez-Araneda, "Towards a holistic model for quality of learning object

repositories: A practical application to the indicator of metadata compliance”, *Emerald Publ. Ltd.*, 2016.

- [17] J. R. Smith y P. Schirling, “Metadata standards roundup”, *IEEE Multimed.*, vol. 13, no. 2, pp. 84–88, 2006.
- [18] O. P. Sharma, “Quality Indicators of Scientific Research”, *Indian J. Microbiol.*, vol. 52, no. 2, pp. 305–306, 2012.
- [19] L. Bornmann and W. Marx, “The h index as a research performance indicator”, *Eur. Sci. Ed.*, vol. 37, no. 3, pp. 77–80, 2011.
- [20] Cornell University Library, “LibGuides: Measuring your research impact: i10-Index”, 2017.