

INSTITUTO TECNOLÓGICO DE ORIZABA
DIVISIÓN DE ESTUDIOS DE POSGRADO E INVESTIGACIÓN
MAESTRÍA EN SISTEMAS COMPUTACIONALES

TESIS:

Módulo de recuperación de metadatos de referencias de artículos

PRESENTADO POR:

Alma Delia Apale Zitzihua M19011611

PARA OBTENER EL GRADO DE:

Maestro en Sistemas Computacionales

DIRECTOR DEL ANTEPROYECTO DE TESIS:

M.R.T. Ignacio López Martínez

Orizaba, Ver a 04 de junio de 2021.

Contenido

Introducción	4
Capítulo 1. Antecedentes	6
1.1 Marco teórico	6
1.1.1 Identificadores de autor, revista y artículo: ORCID, DOI y ISSN.....	6
1.1.2 DOI.....	6
1.1.3 ORCID	7
1.1.4 ISSN	7
1.1.5 Metadatos	8
1.1.6 Crossref	8
1.1.7 Algoritmo de búsqueda	8
1.1.8 Gestores de referencias bibliográficas.....	9
1.1.8.1 EndNote.....	9
1.1.8.2 Mendeley.....	10
1.1.8.3 Zotero	10
1.2 Planteamiento del problema	1
1.3 Objetivo general y específicos	1
1.3.1 Objetivo General	1
1.3.2 Objetivos específicos	2
1.4 Justificación.....	2
Capítulo 2. Estado de la práctica.....	4
2.1 Trabajos relacionados.....	4
2.2 Análisis comparativo.....	8
2.3 Propuesta de solución.....	9
2.3.1 Lenguaje de programación lado del servidor: Node.js.....	10
.....	10
2.3.2 Entorno de desarrollo integrado: NetBeans	11
2.3.3 Sistema Gestor de Base de Datos (SGBD): MariaDB	11
2.3.4 Metodologías de desarrollo de software: Scrum	12
2.3.5 Justificación de la solución	13
Capítulo 3. Aplicación de la metodología.....	14
3.4.2.1 Actividades.....	14

3.4.3 Entregables	24
Capítulo 4. Resultados	25
Título del artículo: A study on LIWC categories for opinion mining in Spanish reviews ..	25
Conclusiones	28
Bibliografía	29
TABLA 1 ANÁLISIS COMPARATIVO DE GESTORES DE REFERENCIAS BIBLIOGRÁFICAS.	0
TABLA 2 ANÁLISIS COMPARATIVO DE TRABAJOS RELACIONADOS.	8
TABLA 3 PROPUESTA SELECCIONADA.....	13
TABLA 4 DESCRIPCIÓN DE REQUERIMIENTOS.	15
FIGURA 1 LOGOTIPO OFICIAL DOI.....	6
FIGURA 2 LOGOTIPO OFICIAL ORCID.	7
FIGURA 3 LOGOTIPO OFICIAL ISSN.	7
FIGURA 4 LOGOTIPO OFICIAL CROSSREF.	8
FIGURA 5 LOGOTIPO OFICIAL ENDNOTE.....	9
FIGURA 6 LOGOTIPO OFICIAL MENDELEY.....	10
FIGURA 7 LOGOTIPO OFICIAL ZOTERO.....	10
FIGURA 8 ARQUITECTURA PRELIMINAR.....	9
FIGURA 9 LOGOTIPO OFICIAL NODE.JS.	10
FIGURA 10 LOGOTIPO OFICIAL NETBEANS.	11
FIGURA 11 LOGOTIPO OFICIAL MARIADB.....	11
FIGURA 12 LOGOTIPO OFICIAL SCRUM.	12
FIGURA 13 DIAGRAMA DEL FUNCIONAMIENTO DE LAS HERRAMIENTAS SELECCIONADAS....	16
FIGURA 14 FUNCIONAMIENTO DEL MÓDULO WEB.....	17
FIGURA 15 MOCKUP INTERFAZ DE REGISTRO DE USUARIOS EN EL MÓDULO WEB.	18
FIGURA 16 MOCKUP INTERFAZ DE INICIO DE SESIÓN.	¡ERROR! MARCADOR NO DEFINIDO.
FIGURA 17 MOCKUP INTERFAZ DE FUNCIONAMIENTO AUTOMATIZADO DEL MÓDULO.....	18
FIGURA 18 BÚSQUEDA DE DOI'S DE ARTÍCULOS.....	19
FIGURA 19 BÚSQUEDA DE DOI'S DE PREFIJO DETERMINADO.	19
FIGURA 20 BÚSQUEDA CON PALABRAS CLAVE.	19
FIGURA 21 API DE CROSSREF RESULTADOS.	20
FIGURA 22 DATOS DE REGISTRO.....	21
FIGURA 23 DATOS DE ACCESO ASIGNADOS POR ORCID.	22
FIGURA 24 CRONOGRAMA DE ACTIVIDADES 2020 - 2021.....	24

Resumen

En la actualidad surge la necesidad por parte de estudiantes e investigadores de extraer e integrar la información bibliográfica de los artículos científicos que han sido publicados en distintas revistas o libros de divulgación científica que se concentran en internet con el fin de visualizar, extraer, almacenar y gestionar estos resultados en los repositorios de sus instituciones que servirán para llevar un control, así como el seguimiento del trabajo y producción de trabajos de investigación. Los metadatos son indispensables al momento de realizar la búsqueda en un sentido referencial de una investigación de un artículo digital, entre los metadatos que contienen las referencias de un artículo uno de los más relevantes que poseen es el *Identificador de Objeto Digital* (DOI); el cual facilita la localización dentro de las bases de datos en las que son almacenados, ya que se compone de un código alfanumérico que es único asignado para cada objeto digital. La cantidad de información concentrada en Internet es enorme y conforme pasa el tiempo se van desarrollando nuevas y mejores herramientas así también métodos de recuperación de esta información. En este trabajo de investigación se presenta el diseño de un módulo Web automatizado para la recuperación referencial de metadatos de los artículos científicos de distintas bases de datos. Adicionalmente se presenta y se describe una arquitectura de los componentes y su funcionamiento.

Introducción

La innovación, así como los nuevos descubrimientos científicos requieren largos y rigurosos periodos de investigación que resultan en nuevas herramientas y métodos que ayudan a mejorar la vida diaria, este tipo de resultados en su mayoría se reflejan actualmente en productos digitales como lo son artículos de divulgación científica. Es importante recordar que estos productos de investigación se clasifican e identifican mediante su referencia bibliográfica; la cual se compone principalmente de un conjunto de metadatos como: autor, título, fecha de publicación, nombre de la revista, así como un identificador único (DOI), estos datos sirven para buscar artículos.

Existen herramientas como los gestores de referencias [1] que se encargan de gestionar grandes cantidades de productos digitales de investigación, además permiten buscar y almacenar de manera organizada dichos productos. Mendeley [2], Zotero [3], EndNote [4] son tres de los gestores más conocidos y utilizados, entre sus características se destacan que Mendeley provee 2 GB de almacenamiento gratuito y funciona mediante una aplicación de escritorio que cuenta con cinco maneras de extracción de artículos los cuales están incorporados a las bases de datos pertenecientes a la editorial Holandesa Elsevier [5], estas herramientas utilizan *plugins* (programa informático de inserción) [6] en páginas Web y bases de datos pero para la búsqueda se requiere ingresar la mayor parte de los datos del trabajo de investigación en algunos casos no se presentan problemas de compatibilidad. Por otro lado, Zotero lleva a cabo la misma función, pero a diferencia del gestor anterior, solo provee de forma gratuita 300 MB de almacenamiento y presenta problemas de compatibilidad con algunas versiones de Word. Finalmente, el gestor EndNote permite almacenar distintos datos de los trabajos de investigación como imágenes y referencias, pero es muy poco conocido por la comunidad investigadora. Los tres gestores descritos anteriormente son de gran utilidad al momento de realizar una investigación, pero existen varios motivos que hacen de ellos muy complejos al momento de utilizarlos como dificultades al momento de la instalación en los dispositivos que en ocasiones presentan incompatibilidad en las versiones de software, el uso de complementos en los distintos navegadores, entre otros. ; en este mismo contexto es importante mencionar la función que llevan a cabo los *Repositorios Institucionales* (RI) que se rigen bajo la iniciativa *Open Access*(OA) el cual consiste en que cualquier investigación que sea financiada con recurso

público debe brindar el libre acceso, pero para ello se estipulan una serie de requisitos dentro de los documentos que pone a disposición el Repositorio Nacional que es operado por el *Consejo Nacional de Ciencia y Tecnología* (CONACYT). A causa de esto es necesario el desarrollo de una herramienta que permita recopilar estos productos de investigación de manera referencial para tener un mayor control de la producción científica y académica de las instituciones y que además sea intuitiva, sencilla y fácil de utilizar para la extracción de referencias con base en las necesidades del usuario final.

Tomando en cuenta lo anterior este trabajo presenta el desarrollo de un módulo Web automatizado que permite recuperar información de referencias de artículos a nivel de metadatos para asegurar una información más precisa y sin duplicidad, es decir, una búsqueda más simple ingresando la menor cantidad de datos del trabajo de investigación y así localizarlo en las principales bases de datos; el capítulo 1 presentan los antecedentes sobre métodos de extracción de metadatos con diferentes tipos de tecnologías; el capítulo 2 se desglosa el estado de la practica; en el capítulo 3 se realizó la aplicación de la tecnología, en seguimiento en el capítulo 4 se presentan los resultados y finalmente en el capítulo 4 se redactaron las conclusiones y recomendaciones referentes a este trabajo de investigación.

Capítulo 1. Antecedentes

1.1 Marco teórico

En este capítulo se presentan los términos más importantes que se utilizan en la gestión de trabajos académicos en las distintas plataformas existentes, en la actualidad existen muchas herramientas en la red que sirven para gestionar los productos de investigación y estas, a su vez, cuentan con un propio método de identificación, es necesario enlistar en un principio los tipos de identificadores por formato y posteriormente los conceptos relevantes sobre tecnologías a utilizar.

1.1.1 Identificadores de autor, revista y artículo: ORCID, DOI y ISSN

Actualmente existen identificadores propios de cada tipo de producto generado durante o después de una investigación con el fin de tener un mejor control de temas y autores que las realizaron, a continuación, se enlistan estos identificadores:

1.1.2 DOI



Figura 1 Logotipo oficial DOI.

El sistema de identificación DOI (“*Digital Object Identifier*”, Identificación de Material Digital), es una infraestructura estandarizada para la identificación única y persistente de los diferentes tipos de objetos [7]. Este sistema de identificación está conformado por dos componentes: el primero un prefijo que identifica la editorial de procedencia utilizando un número, y el segundo un sufijo el cual identifica el objeto digital en este caso el artículo, ambos conforman la estructura de un DOI y este funciona de manera general como una huella digital de los productos de una investigación esto a causa de que se le asignan datos únicos

en su registro. Otra de las ventajas del uso de este identificador es que permite consultar más rápido en distintas bases de datos, al igual que se asegura la propiedad intelectual del documento.

1.1.3 ORCID



Figura 2 Logotipo oficial ORCID.

ORCID (*“Open Researcher and Contributor ID”*, Identificador Abierto de Investigador y Colaborador), es una organización sin fines de lucro que se encarga de proveer un sistema de identificación única a los distintos formatos resultado de la investigación [8], o en su caso la innovación para ser utilizados de forma más segura y exista una relación más fácil de identificar.

1.1.4 ISSN

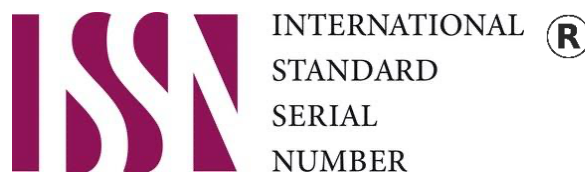


Figura 3 Logotipo oficial ISSN.

El número ISSN (*“International Standard Serial Number”*, Número Internacional Normalizado de Publicaciones Seriadas), este identificador está compuesto por ocho dígitos que sirven para identificar publicaciones periódicas y editadas, productos digitales o físicos. El ISSN está avalado en más de 60 países con centros especializados en cada uno de ellos, incluyendo España y Latinoamérica donde se asigna oficialmente un número único e irremplazable a las revistas y colecciones seriadas [9].

1.1.5 Metadatos

La bibliografía está integrada por un conjunto de metadatos, pero entendamos el término metadatos: La primera acepción que se le dio (y actualmente la más extendida) fue la de dato sobre el dato, ya que proporcionaban la información mínima necesaria para identificar un recurso; en este mismo concepto se afirma que incluye información descriptiva sobre el contexto, calidad y condición o características del dato [10]. Sin embargo, si se habla de los metadatos en el contexto de formatos de investigación como es el caso, la bibliografía, así como los distintos tipos de identificadores forman un conjunto de metadatos porque en ellos contienen distintos datos almacenados. Para el proceso de extracción de dichos metadatos se requiere una limpieza y una catalogación en un orden específico de manera que sea lo que espera obtener el usuario al solicitar la búsqueda.

1.1.6 Crossref



Figura 4 Logotipo oficial Crossref.

Es una agencia oficial de registro de identificadores de objetos digitales (DOI) de la Fundación Internacional DOI. Está dirigido por *Publishers International Linking Association Inc.* (PILA) y se lanzó a principios de 2000 como un esfuerzo cooperativo entre los editores para permitir la vinculación de citas cruzadas entre editores persistentes en revistas académicas en línea [11]. Cabe mencionar que la asignación de los identificadores tiene un costo dependiendo las publicaciones y se funge como editor para asignar estos a los trabajos de investigación.

1.1.7 Algoritmo de búsqueda

Es aquel que está diseñado para localizar un elemento con ciertas propiedades dentro de una estructura de datos; por ejemplo, ubicar el registro correspondiente a cierta persona en una

base de datos, o el mejor movimiento en una partida de ajedrez [12]. Este término se menciona ya que posteriormente se hará la selección de un algoritmo que permita realizar de mejor manera la extracción de metadatos para una búsqueda más precisa.

1.1.8 Gestores de referencias bibliográficas

Son programas que permiten crear, mantener, organizar y dar forma a referencias bibliográficas de artículos de revistas o libros; obtenidas automáticamente de una o de varias fuentes de información (bases de datos, revistas, páginas web...), y que añaden a esta función básica su versatilidad para generar cientos de formatos de entrada y salida, utilizados para citar referencias bibliográficas en los trabajos de investigación. Esto es fundamentalmente lo que les diferencia de cualquier otro programa de bases de datos, por una parte la capacidad para compilar información desde la mayoría de las fuentes de información, y por otra la integración en los programas de proceso de textos (*Word, Open Office*) para facilitar a los investigadores la inclusión de citas en los trabajos de investigación [1].

En este apartado se presentan los 4 gestores de referencias más utilizados para la gestión de documentos científicos:

1.1.8.1 EndNote



Figura 5 Logotipo oficial EndNote.

Es un programa de gestión de referencias bibliográficas, diseñado para agilizar la redacción de trabajos de investigación. Con *EndNote*, se logra: Recopilar, de una forma rápida y sencilla, la información de referencia procedente de una amplia variedad de fuentes de datos en línea, como, por ejemplo, *PubMed*, *Google Scholar* y *Web of Science*. Almacenar las referencias en su propia biblioteca protegida con contraseña y a la que se accede desde cualquier sitio en el que disponga de acceso a Internet. Compartir referencias con otros usuarios de *EndNote* para facilitar la colaboración. Dar formato de cita al texto en Microsoft

Word con la herramienta *Cite While You Write* (Citar mientras escribe), permite introducir referencias y aplicar formato a los trabajos al instante [4].

1.1.8.2 Mendeley



Figura 6 Logotipo oficial Mendeley.

Es un gestor de referencias gratuito y red social académica con sede en Londres creado por Jan Reichelt, Victor Henning y Paul Foeckler, que fuera adquirido en abril de 2013 por Elsevier. Mendeley permite generar automáticamente bibliografías. Colaborar fácilmente con otros investigadores en línea. Importar fácilmente documentos de otro software de investigación. Localizar los documentos pertinentes de acuerdo con lo que está leyendo. Leer artículos con aplicación para iPhone. Acceder a sus documentos en línea desde cualquier lugar [2].

1.1.8.3 Zotero



Figura 7 Logotipo oficial Zotero.

Es un gestor de referencias bibliográficas, de acceso libre y gratuito desarrollado por el “Roy Rosenzweig Center for History and New Media” de la Universidad George Mason. Trabaja como una extensión del navegador Firefox que permite a los usuarios crear su propia base de datos (biblioteca) y llevar a cabo acciones como: guardar, manejar, buscar, organizar y citar fuentes de todo tipo. Zotero reside en el navegador Firefox, aunque la versión Standalone Alpha (desktop) funciona con los navegadores Safari y Chrome mediante un *plugin* [3].

Repositorio Institucional (RI)

Es un archivo digital que realiza la función de recopilar, administrar y difundir los productos resultantes de una investigación científica o académica. Se estipula en el documento *Lineamientos Generales* en el Repositorio Nacional que una investigación financiada con recursos públicos debe estar disponible al público en general, además se resalta la importancia de dar a conocer los avances de una investigación para mostrar el impacto de la producción científica a nivel nacional e internacional, la difusión puede lograr retroalimentación por más expertos en el tema, así como facilitar el acceso a esta información a quienes les sea útil para derivar otra investigación, debido a esto, el contenido de este archivo digital debe ser de calidad, actualizarse constantemente, seguro y por supuesto cumplir con los aspectos legales como el respeto a los derechos de propiedad intelectual.

Evaluación por pares

De acuerdo con las definiciones expuestas en el capítulo tercero del documento LER (Lineamientos específicos para repositorios) *Capítulo I Disposiciones Generales*, la evaluación por pares es un proceso que tiene como objetivo validar la calidad, originalidad, factibilidad, rigor científico y demás aspectos de importancia de los productos que serán almacenados en el RI.

Acceso Abierto

El acceso abierto (en inglés, Open Access, OA) es el acceso gratuito a la información y al uso sin restricciones de los recursos digitales por parte de todas las personas. (*¿Qué es acceso abierto?* | UNESCO, 2021). La editorial holandesa Elsevier define Acceso Abierto o mejor conocido por sus siglas en inglés como “OA, Open Access” como un acontecimiento que surgió de manera simultánea con las TICs (Tecnologías de la información) y se traduce como el conocimiento abierto que se realiza a través de internet, con esta iniciativa son beneficiados especialmente estudiantes, académicos y científicos en el desarrollo de alguna investigación. Es de gran importancia mencionar que esta *Iniciativa Open Access de Budapest* surgió en el año 2002 por la Open Budapest Foundation la cual contempló algunas recomendaciones para favorecer el acceso a esta información al público, posterior a esto se redactaron dos iniciativas más: *Bethesda* y *Berlín* las cuales actualmente están en funcionamiento y aplicándose en este contexto de ciencia abierta, otro aspecto importante es la clasificación actual que divide en dos rutas de acceso abierto: la vía dorada (Golden Open Access) y la vía verde (Green Open Access).

“El Golden Open Access o vía dorada supone para el lector el acceso público y gratuito al artículo final (esto es, tal y como se ha publicado), siendo este acceso inmediato y permanente, que es el proceso por el que el autor (que ha publicado su artículo en una revista

de suscripción) deposita su artículo, una vez aceptado (postprint), en una web o repositorio de recursos digitales, sin necesidad de pagar el APC (los costes se cubren con las suscripciones)”

En la guía para la evaluación de repositorios institucionales de investigación se exponen las siguientes ventajas del OA:

La investigación se hace más efectiva y sus resultados más visibles cuando se eliminan las barreras legales, comerciales y tecnológicas de acceso a la información científica. Además, el acceso abierto previene la duplicación, fomenta la transferencia tecnológica y de conocimiento y promueve la innovación. Todos los actores implicados en el sistema de comunicación de la información científica se benefician del acceso abierto a la investigación y a los datos generados por ésta.

A continuación se muestra una tabla comparativa de los gestores de referencias antes descritos en la cual se exponen sus características generales ventajas y desventajas.

Principales características de los gestores bibliográficos:

Características	EndNote	Zotero	Mendeley
Costo	Gratuito usuarios UD	Gratuito / Cuenta Premium	Gratuito / Cuenta Premium
Tipo de aplicación	WEB	WEB/Escritorio	WEB /Escritorio
Almacenamiento	2 GB	300 MB (cuenta gratuita)	2 GB (cuenta gratuita)
Copia de seguridad	NO	SÍ	SÍ

Tabla 1 Análisis comparativo de gestores de referencias bibliográficas.

1.2 Planteamiento del problema

Actualmente vivimos en un entorno globalizado y en él la ciencia es causante de los resultados de largos periodos de investigación que han aportado lo que existen en descubrimientos e innovación como herramientas que sirven para mejorar la vida diaria para el ahorro de los recursos en la realización de distintas actividades.

Es entonces que en este punto de la actualidad los productos resultantes son clasificados e identificados con ayuda de la bibliografía la cual guarda un conjunto de datos como: título, fecha de publicación, autores, páginas, entre otros, la cual es de gran ayuda para la clasificación y gestión de los distintos productos en sus distintos formatos que existen para cada uno específicamente.

Teniendo en cuenta que no se ha realizado un sistema que permita una búsqueda y obtención de datos más precisos los actualmente existentes arrojan datos poco precisos y ligeramente diferentes, es importante mencionar que se realiza mediante un proceso a nivel de limpieza y búsqueda de metadatos. Esto representa una gran problemática para acceder hoy en día a la ciencia de la información en una forma más rápida y eficaz en la búsqueda y la necesidad de administrar los trabajos de investigación para dar seguimiento a futuros resultados, aportaciones en otras investigaciones tanto personalmente o como parte de una organización que quiera conocer o llevar el control de sus investigadores y su estatus actual dentro de las bases de datos científicas.

1.3 Objetivo general y específicos

1.3.1 Objetivo General

Desarrollar un módulo que permita automatizar e integrar la recuperación de referencias en diferentes formatos de bases de datos científicas y utilizando distintos métodos de búsqueda.

1.3.2 Objetivos específicos

- Recopilar información sobre trabajos similares existentes para hacer una comparación entre ellos.
- Estudiar y analizar el estado del arte de los métodos de recuperación de referencias bibliográficas para seleccionar el adecuado.
- Analizar los formatos de referencias de artículos y bases de datos además conocer su estructura.
- Investigar los estándares y protocolos aplicables a este desarrollo
- Investigar bases de datos existentes del ámbito de divulgación científica para la extracción de metadatos.
- Seleccionar una metodología adecuada al desarrollo del módulo.
- Determinar las tecnologías que se utilizarán para la implementación del producto a desarrollar.
- Diseñar la arquitectura del funcionamiento del módulo para iniciar el desarrollo adecuadamente.
- Implementar el módulo con base en la arquitectura diseñada.
- Validar el método implementado por medio de un caso de estudio.

1.4 Justificación

Se realizó una investigación y se localizaron tecnologías que utilizan los existentes gestores de referencias bibliográficas y que funcionan para recopilación de información bibliográfica de trabajos de investigación y que son inconsistentes, poco confiables, además muestran duplicidad y con mala conexión a nivel de metadatos así como la escasa utilización de los identificadores internacionales de los distintos formatos para revistas, artículos y libros y los estándares que regulan el manejo de este tipo de información, también es importante retomar el funcionamiento de los repositorios institucionales (RI) pero que dependen de la decisión de cada autor en depositar sus trabajos de investigación de manera voluntaria en estos sistemas de información.

Después observar y analizar la información que proveen estos dos sistemas se concluye que es necesario desarrollar un recopilador que resuelva las problemáticas antes mencionadas, con el fin de proveer a los investigadores una herramienta eficaz y eficiente que permita la recopilación de

trabajos en un sentido referencial para llevar el control de las publicaciones en las distintas bases de datos, así como monitorear el estatus de la información que se incluye en otras investigaciones. Por medio de este trabajo se verán beneficiados estudiantes, docentes, investigadores, así como organizaciones que necesiten acceder a esta información de manera general y grupal de forma más sencilla.

Capítulo 2. Estado de la práctica

En la actualidad la información en la web sobre artículos científicos es enorme y conforme pasa el tiempo se van desarrollando nuevas herramientas y métodos de recuperación de la información. Para contribuir en la mejora e innovación en este aspecto, es necesario conocer los trabajos relacionados y las herramientas de software existentes. Es por ello que a continuación se presenta una recopilación de trabajos similares y posteriormente con el apoyo de una tabla se describen los datos más relevantes de cada trabajo. Finalmente se presenta una conclusión acerca de esta investigación realizada.

2.1 Trabajos relacionados

La producción en el ámbito de divulgación científica crece día a día y genera que los investigadores, así como las instituciones donde llevan a cabo este proceso se vean en la necesidad de llevar el control de manera organizada, así como de cada uno de sus productos generados; se han desarrollado herramientas que ponen a disposición para buscar, citar y almacenar estos trabajos de investigación. A continuación, se presenta una revisión del estado del arte de los trabajos existentes que han utilizado herramientas y/o métodos de búsqueda y extracción de datos de artículos científicos.

Klein y Van de Sompel [13] afirmaron que en las últimas dos décadas, la comunicación de la investigación ha pasado de ser un esfuerzo basado en el papel a una empresa digital basada en la Web. Con el paso del tiempo el proceso de investigación ha comenzado a evolucionar, ya que pasó a ser una actividad abierta y visible a nivel mundial. Con el fin de apoyar a los distintos grupos de investigadores surgió una gran variedad de repositorios, bases de datos, plataformas virtuales como una herramienta de divulgación científica. Sin embargo, algo que ha pasado rara vez pero que significa mucho para los usuarios es que las plataformas desaparecen sin dejar rastro causando la pérdida de información importante.

Algunas de ellas no ofrecen garantías del contenido publicado, es por eso que se propuso un nuevo paradigma de archivo que se centra en el descubrimiento de entidades y artefactos web en el ámbito del enfoque de archivo Web. Dado que el ORCID surgió como una infraestructura Web académica de alto potencial que asigna identificadores únicos digitales a los docentes, permite listar identidades Web adicionales, así como artefactos. Sin embargo, también se descubrió que los crecimientos de ORCID tienen una tasa muy significativa que supera el crecimiento de los investigadores y eso ha causado también que varias plataformas últimamente opten por añadir esta función a sus tecnologías. Por lo tanto, existe una posibilidad real de que ORCID alcance un nivel de cobertura en un futuro próximo más adecuado para resolver las necesidades de los investigadores.

Holzmann y Runnwerth [14] mencionaron que existen muchos tipos de identificadores para referirse a los trabajos académicos principalmente hicieron notar la existencia de DOI (*Digital Object Identifier*, Identificador de Objetos Digitales) Sistema que se inició en el año 1998 [7] por Crossref una organización de miembros sin ánimos de lucro [11] , estos identificadores están constituidos por datos en conjunto, referentes a la obra y que son llamados metadatos. Como problemática presentada en este artículo se argumentó que las referencias provenientes de sitios Web como lo son los blogs o programas informáticos, son muy vagas, aunque esta contenga la URL (*Uniform Resource Locator*, Identificador de recursos uniforme) [15] y la fecha en que se visitó, ya que, los sitios están en constante modificación de su contenido. Por lo anterior se presentaron los Micro Archivos que son colecciones microscópicas como su nombre lo dice, de los recursos archivados en la Web, utilizados para describir los objetos o entidades mediante el prototipo Micrawler (*Micro Crawler*, micro rastreador) que es una implementación y prueba de referencia encargado de la gestión de estos conjuntos de micro información. Micrawler tiene como objetivo principal que los archivos consultados en la Web archiven todos los recursos relacionados incluyendo el código fuente resultando en un micro Archivo para una consulta permanente mediante una URL corta o un mediante DOI.

A su vez Bangert y Frances [16] afirmaron que nos encontramos en un panorama en el que los productos académicos aumentan exponencialmente, a causa de esto; los Identificadores Persistentes (PID) son una tecnología clave para permitir el acceso y la interoperabilidad entre los sistemas involucrados en la comunicación académica, estos identificadores permiten rastrear y visualizar un trabajo de investigación, como resultado del surgimiento de nuevas plataformas que almacenan estos documentos de investigación surgió la necesidad de implementar mejoras en esta tecnología, es así, que varias organizaciones trabajan para mejorar la integración en la infraestructura de investigación internacional y se propuso que sea a través de proyectos de colaboración con plataformas como ORCID que provee una forma de identificación al autor y la Red de Interoperabilidad de DataCite.

Este proyecto consistió en el diseño y la entrega de los servicios de PID en la Biblioteca de la UNSW que está guiado por las características de los identificadores como es el caso de DOI.

Los identificadores asignados a los resultados de la investigación son interoperables, basados en las fuentes verídicas, y contienen metadatos legibles por humanos y máquinas. DOI permite tener un acceso persistente al recurso está garantizado por la biblioteca como custodio de los identificadores y del contenido del repositorio asociado. Los resultados fueron que a medida que los identificadores pasan a formar parte de cada etapa del ciclo de vida de la investigación, el desafío para las instituciones será seguir rigiendo su asignación de manera efectiva, seguir las normas de la comunidad, y optimizar su uso para y por los investigadores.

Además Yang y Zhang [17] dedujeron en vista de que diversas plataformas albergan miles de trabajos de investigación científica los investigadores tienen un acceso ilimitado a estas herramientas digitales, existen distintas formas de realizar la búsqueda de información y una de las más utilizadas es empleando las palabras claves para una recopilación precisa y necesaria de lo que se busca, esto se lleva a cabo por medio de la utilización de algoritmos de búsqueda o métodos de rastreo, en algunos casos estas formas muestran artículos más citados recientemente o su reputación dependiendo el autor, esto en algunos casos no es lo que se requiere para el trabajo de investigación que se está desarrollando. Teniendo esta problemática se propuso la utilización del

algoritmo de integración de texto llamado "TextRank" que es utilizado para ayudar a los investigadores en la realización de la revisión de la literatura. El objetivo principal de TextRank es resumir una obra determinada de documento que se encuentre contenido en alguna base de datos, por lo que se dice de acuerdo a su funcionamiento es que este algoritmo se basa en la minería de texto y al ser compleja se requiere un gran esfuerzo humano.

A continuación, se presenta una tabla comparativa de trabajos relacionados y el alcance que cada uno tiene respecto a los métodos y tecnologías utilizadas.

Tabla I. Comparativa de trabajos que muestran la falta de un sistema Web que se complemente con estas tecnologías, en el que se incluyan identificadores de autores, así como de artículos para recopilar y gestionar información precisa de obras de investigación que se encuentren almacenadas en distintas bases de datos y de diferentes editoriales

2.2 Análisis comparativo

Artículo	Problema	Contribución	Tecnologías	Resultado	Estado
Holzmann Y Runnwerth [14]	Los métodos de recolección de identificadores únicos respaldados por conjuntos de meta información existentes tienen graves desventajas y no siempre son adecuados además de que no son muy precisos y no garantizan la persistencia.	Micro Archivos como ricas representaciones de objetos digitales, que conectan semántica y lógicamente los recursos archivados.	Micrawler Micro Archivos	Aumento en las cifras de conceptos novedosos que permitió citas, coherentes se observó mucho potencial en tales colecciones microscópicas estableciendo la semántica que falta y un enlace lógico entre los recursos de la web combinado con una incrustación temporal.	Terminado
Bangert y Frances [16]	Surgió la necesidad de implementar mejoras en esta tecnología es así que varias organizaciones han trabajado para mejorar la integración en la infraestructura de investigación internacional.	Se propuso mejorar estudiando proyectos de colaboración como el ORCID y la Red de Interoperabilidad de DataCit	PID	Los resultados fueron que a medida que los identificadores pasan a formar parte de cada etapa del ciclo de vida de la investigación, el desafío para las instituciones será seguir rigiendo su asignación de manera efectiva, seguir las normas de la comunidad, y optimizar su uso para y por investigadores.	Desarrollo
Yang & Zhang [17]	La información a la que un investigador accede convenientemente es ilimitada entonces es virtualmente imposible para uno seguir la gran cantidad de desarrollos que se están creando en el día a día. Además de que la literatura basada en la clasificación no proporciona a los investigadores las publicaciones más recientes, limitando así la eficiencia y la efectividad de la revisión de la literatura.	Se propuso el algoritmo de integración de texto llamado "TextRank" que es utilizado para ayudar a los investigadores en la realización de la revisión de la literatura.	TextRank	En comparación con el resumen oficial de un artículo, el resumen basado en el texto es más flexible (permite controlar el número de frases) y más objetiva (el resumen se basa en la semántica del artículo).	Desarrollo
Klein Y Van de Sompel [13]	La historia ha demostrado que incluso las plataformas web más populares desaparecen sin dejar rastro. Para empeorar las cosas, rara vez ofrecen garantías de archivo explícitas; muchas veces son bastante contrapuestas.	Se propuso un nuevo paradigma de archivo que está dirigido a los huérfanos académicos basados en la web. El primer y segundo paso de este paradigma se centra en el descubrimiento de entidades y artefactos web en el ámbito de nuestro enfoque de archivo web.	ORCID	Se descubrió que los crecimientos de ORCID tienen una tasa muy significativa que supera el crecimiento de los investigadores. Por lo tanto existe una posibilidad real de que ORCID alcance un nivel de cobertura en un futuro próximo más adecuado para resolver las necesidades de los investigadores.	Desarrollo

Tabla 2 Análisis comparativo de trabajos relacionados.

En la tabla 1 se muestra un análisis comparativo de los trabajos, para que se observen las diferencias y las similitudes entre ello

2.3 Propuesta de solución

Este capítulo tiene como objetivo presentar una descripción general de la propuesta de solución y su arquitectura preliminar para el funcionamiento del módulo a desarrollar.

La propuesta de solución que se describe consiste en un módulo web que interactúe con distintas bases de datos científicas de los cuales mediante la extracción de metadatos que pertenezcan a trabajos de investigación específicamente de artículos para de ellos recuperar referencias bibliográficas de los autores.

Para interactuar con esta información es necesario hacer uso de las APIs (Interfaces de Programación de Aplicaciones del Navegador, *Application Programming Interface*) a Bases de datos científicas. Es menester resaltar la importancia de los identificadores únicos que poseen actualmente los artículos ya que facilitan su localización dentro de la web.

En la Figura 3.1 se muestra la arquitectura preliminar del módulo a desarrollar.

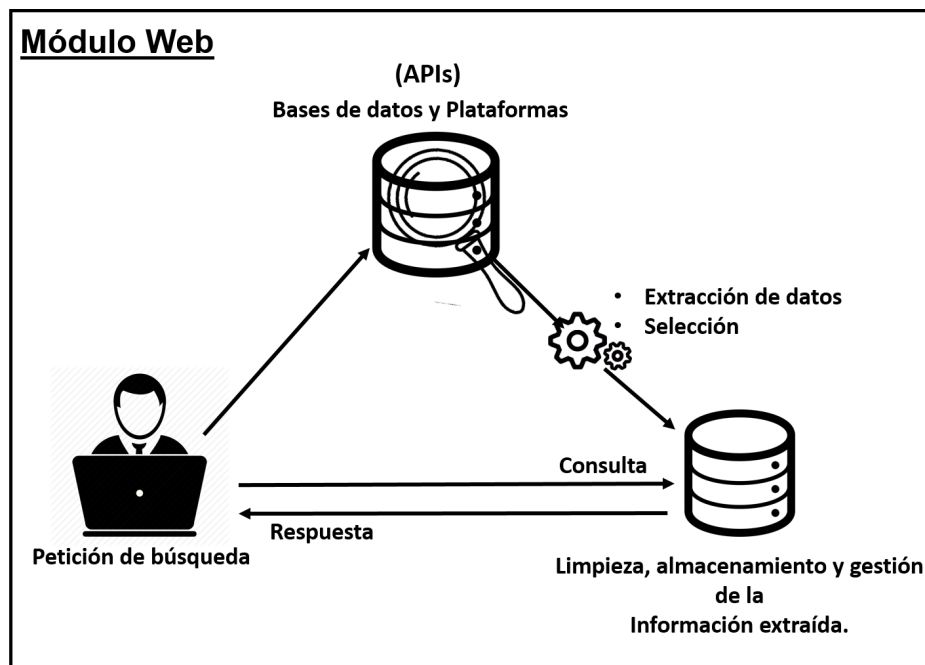


Figura 8 Arquitectura preliminar.

2.3.1 Lenguaje de programación lado del servidor: Node.js



Figura 9 Logotipo oficial node.js.

Node.js es un entorno de tiempo de ejecución JavaScript multiplataforma gratuito y de código abierto que permite a los desarrolladores escribir herramientas de línea de comandos y scripts del lado del servidor fuera de un navegador.

Node.js ejecuta el motor V8 JavaScript, el núcleo de Google Chrome, fuera del navegador. Esto permite que Node.js logre ser muy eficiente.

Una aplicación Node.js se ejecuta en un solo proceso, sin crear un nuevo hilo para cada solicitud. Node.js proporciona un conjunto de primitivas de E / S asíncronas en su biblioteca estándar que evitan que el código JavaScript se bloquee y, en general, las bibliotecas en Node.js se escriben utilizando paradigmas de no bloqueo, lo que hace que el comportamiento de bloqueo sea la excepción y no la norma. Cuando Node.js realiza una operación de E / S, como leer desde la red, acceder a una base de datos o al sistema de archivos, en lugar de bloquear el hilo y desperdiciar ciclos de CPU en espera, Node.js reanudará las operaciones cuando vuelva la respuesta.

Esto permite que Node.js maneje miles de conexiones concurrentes con un solo servidor sin introducir la carga de administrar la concurrencia de subprocesos, lo que podría ser una fuente importante de errores. Este lenguaje tiene una ventaja única porque millones de desarrolladores frontend que escriben JavaScript para el navegador escriben el código del lado del servidor además del código del lado del cliente sin la necesidad de aprender un idioma completamente diferente [18].

2.3.2 Entorno de desarrollo integrado: NetBeans



Figura 10 Logotipo oficial NetBeans.

Es un entorno de desarrollo integrado, modular, de base estándar (normalizado), escrito en el lenguaje de programación Java. El proyecto NetBeans consiste en un IDE de código abierto y una plataforma de aplicación, las cuales es posible usarlas como una estructura de soporte general (*framework*) para compilar cualquier tipo de aplicación.

El lanzamiento de NetBeans IDE 8.2 provee varias características y mejoras nuevas, tales como mejores características de edición JavaScript, soporte para usar estructuras Spring de soporte web, integración MySQL más ajustada y una mejor manera de compartir bibliotecas entre proyectos dependientes. El aclamado soporte para Ruby/JRuby ha mejorado con un nuevo editor de soluciones rápidas (Quick Fix), un administrador para la plataforma Ruby, soporte para depuración rápida en JRuby y muchas otras nuevas características y soluciones [19] .

2.3.3 Sistema Gestor de Base de Datos (SGBD): MariaDB



Figura 11 Logotipo oficial MariaDB.

Este SGBD es una derivación de MySQL que cuenta con la mayoría de características de este e incluye varias extensiones.

Nace a partir de la adquisición de MySQL por parte de Oracle para seguir la filosofía Open Source y tiene la ventaja de que es totalmente compatible con MySQL.

Entre las principales características de este Sistema Gestor de Bases de datos se encuentran:

- Aumento de motores de almacenamiento
- Gran escalabilidad
- Seguridad y rapidez en transacciones
- Extensiones y nuevas características relacionadas con su aplicación para Bases de datos NoSQL.

No tiene desventajas muy aparentes salvo algunas pequeñas incompatibilidades en la migración de MariaDB y MySQL o pequeños atrasos en la liberación de versiones estables [20].

2.3.4 Metodologías de desarrollo de software: Scrum



Figura 12 Logotipo oficial Scrum.

Es un proceso en el que se aplican de manera regular un conjunto de buenas prácticas para trabajar colaborativamente, en equipo, y obtener el mejor resultado posible de un proyecto. Estas prácticas se apoyan unas a otras y su selección tiene origen en un estudio de la manera de trabajar de equipos altamente productivos.

En Scrum se realizan entregas parciales y regulares del producto final, priorizadas por el beneficio que aportan al receptor del proyecto. Por ello, Scrum está especialmente indicado para proyectos

en entornos complejos, donde se necesita obtener resultados pronto, donde los requisitos son cambiantes o poco definidos, donde la innovación, la competitividad, la flexibilidad y la productividad son fundamentales [21].

2.3.5 Justificación de la solución

En esta tabla se muestra la solución propuesta la cual considera la utilización del lenguaje de programación Node.js esto tomando en cuenta sus características se concluye a que es adecuado para el desarrollo web, como entorno de desarrollo se contempló el uso de Netbeans ya que provee soporte para las aplicaciones orientadas a servicios. En cuanto al gestor de base de datos MariaDB por sus características más vanguardistas y flexibilidad y finalmente la metodología Scrum la cual maneja un entorno colaborativo, organizado por iteraciones entre otras cualidades.

En esta tabla se muestra la propuesta seleccionada para el desarrollo:

Propuesta	Lenguaje de programación	IDE	SGBD	Metodología
3	Node.js	NetBeans	MariaDB	Scrum

Tabla 3 Propuesta seleccionada.

Capítulo 3. Aplicación de la metodología

En este capítulo se desglosa la implementación de la metodología Scrum que utiliza un ciclo de vida iterativo para el desarrollo del trabajo de investigación que lleva por nombre: Módulo de extracción de referencias de artículos. Para un mejor control y administración del tiempo para el cumplimiento de cada una de las fases se presenta la gestión de tareas, los requerimientos, así como el seguimiento de cada actividad. Se incluyen en este apartado los documentos propios de la metodología: funcionalidades del proyecto, las tareas que llevarán a cabo, así como el seguimiento que representa el estado actual en el que se encuentran cada una de ellas.

El proceso

El proceso parte de la lista de objetivos/requisitos priorizada del producto, que actúa como plan del proyecto. En esta lista el cliente prioriza los objetivos balanceando el valor que le aportan respecto a su costo (que el equipo estima considerando la definición de hecho) y quedan repartidos en iteraciones y entregas.

Las actividades que se llevan a cabo en Scrum son las siguientes:

- Planificación de la iteración
- Ejecución de la iteración
- Inspección y adaptación

3.4.2.1 Actividades

Artefactos

1. Pila de producto o Product Backlog
2. Pila de sprint o Sprint Backlog

Pila de producto o Product Backlog:

En este artefacto se incluyen todos y cada uno de los requerimientos funcionales del módulo:

Item	Descripción del requerimiento (Backlog)	Responsable
1	Registrar usuarios mediante un identificador único	Alma Delia Apale Zitzihua
2	Obtener ORCID ID autenticado	
3	API de la agencia de DOIs Crossref	
4	API pública de ORCID	
5	Implementar base de datos	
6	Almacenar identificadores de objetos	
7	Usar tokens de acceso actualizados	
8	Almacenar los registros de acceso	
9	Contacto con soporte técnico	

Tabla 4 Descripción de requerimientos.

Búsqueda y recolección de metadatos. Se describe el módulo Web que interactúe con las bases de datos científicas que pertenezcan a trabajos de investigación en sus distintos formatos se recuperen metadatos de referencias bibliográficas de artículos utilizando principalmente los identificadores internacionales y el ingreso de la menor cantidad de datos para el rastreo de los trabajos de investigación, para la búsqueda y extracción.

Selección de herramientas. Para la selección de las herramientas a implementar en el módulo se basó en el siguiente diagrama que muestra el funcionamiento de recolección de productos de investigación:

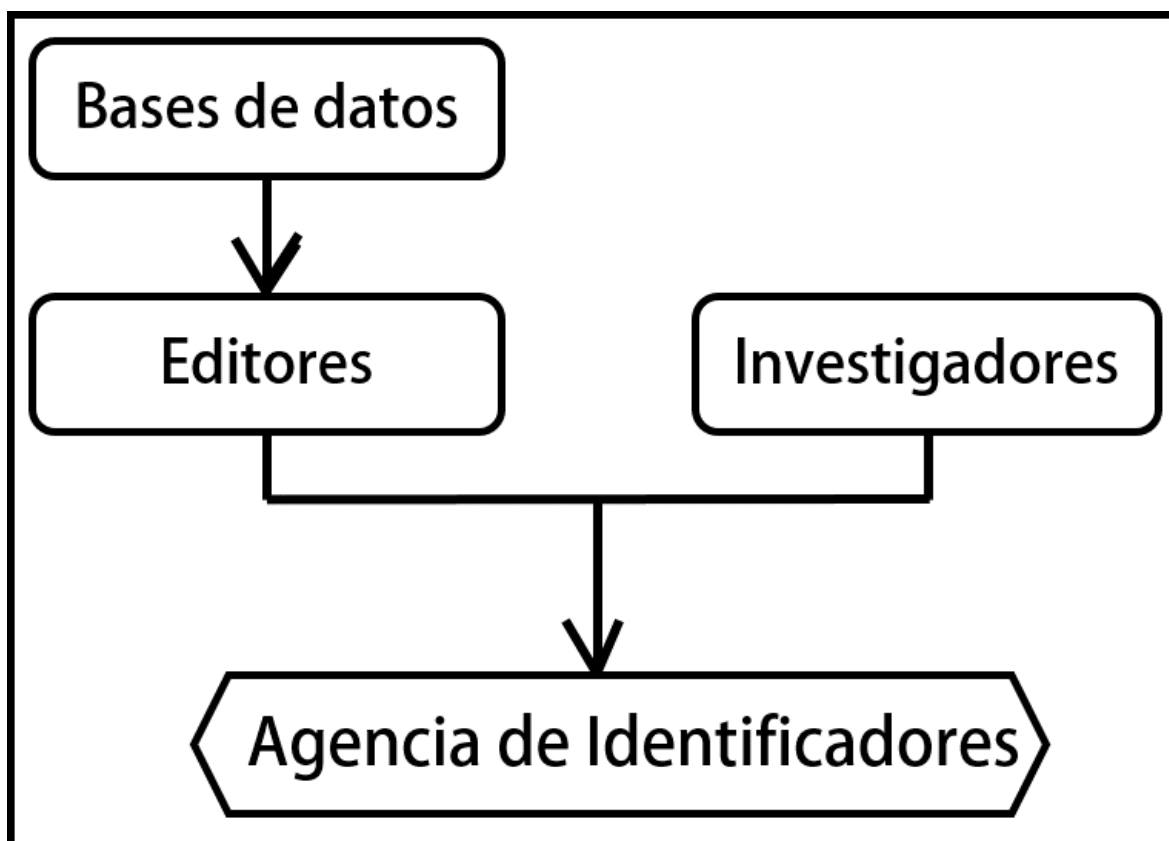


Figura 13 Diagrama del funcionamiento de las herramientas seleccionadas.

Funcionamiento. En primer lugar, las bases de datos como: Scopus, Web of Science, entre otros; pertenecen a la mayor editorial de contenido científico que es Elsevier la cual posteriormente se encarga de depositar sus metadatos a la agencia de identificadores que en este caso es Crossref ; de igual forma los investigadores depositan sus metadatos mediante el uso de ORCID. En base a este funcionamiento se eligió utilizar ambas herramientas para realizar la búsqueda y extracción requerida para el módulo tomando en cuenta la forma en que ellos recolectan la información que se necesita para alimentar el módulo y aprovechando la oportunidad de alcance a esta concentración de datos al público en general, así como desarrolladores y este caso resolviendo los problemas anteriormente planteados.

En el siguiente diagrama se muestra el funcionamiento entre usuario y las herramientas seleccionadas. En primera instancia el usuario realiza una petición de búsqueda de metadatos mediante el módulo que con ayuda de las API's (Interfaces de Programación de Aplicaciones

del Navegador, *Application Programming Interface*) [22] realizará la entrega de la respuesta de la información requerida.

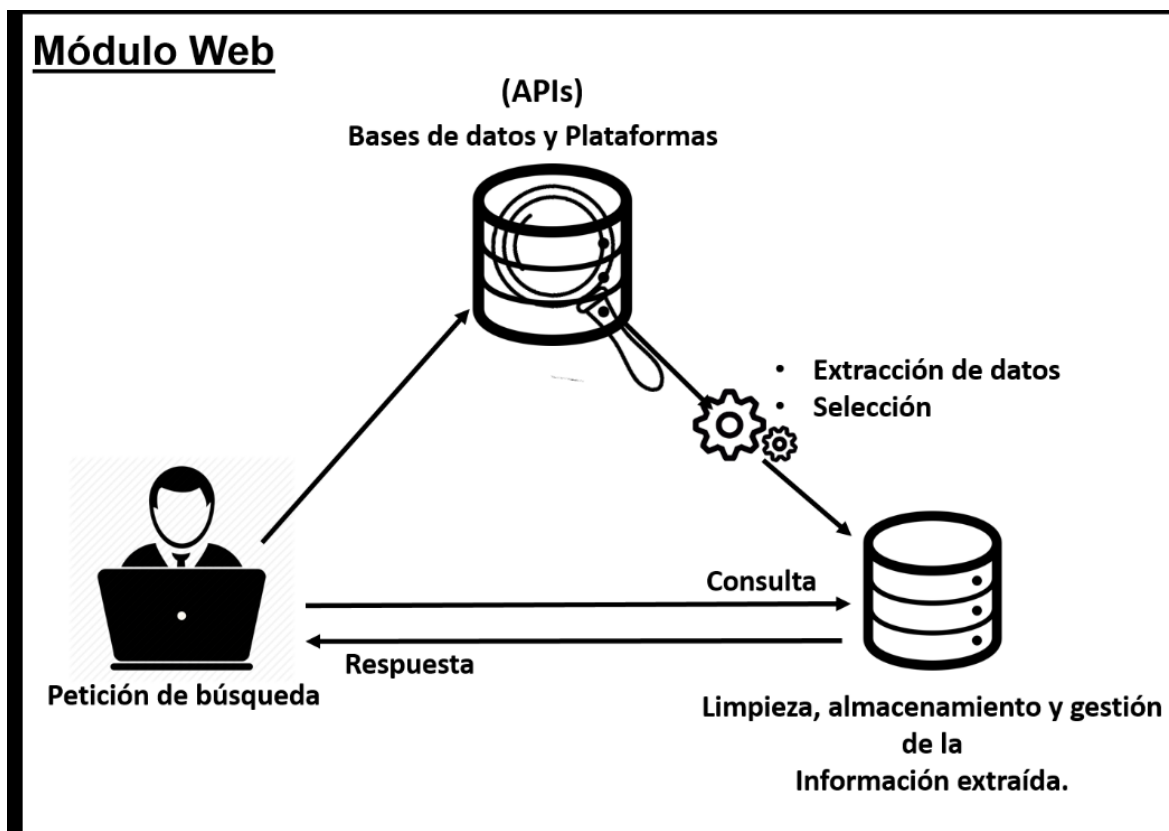


Figura 14 Funcionamiento del módulo Web.

La figura que se presenta a continuación muestra la interfaz que va interactuar con el usuario (estudiante o investigador) que tiene como objetivo realizar la búsqueda automatizada minimizando el ingreso de datos de manera que ingresando el Identificador de Objeto o Identificador de autor puede ser buscado también en libros y revistas. La búsqueda entre las bases de datos será opcional entre utilizar una en específico o ambas simultáneamente. Una vez finalizada la búsqueda el módulo mostrará los resultados obtenidos para la selección de interés por parte del usuario y con las opciones de importación y exportación de sus datos contemplando las políticas de seguridad que pongan en vigor ambas organizaciones que se incluyen en el módulo

Figura 15 Mockup interfaz de registro de usuarios en el módulo Web.

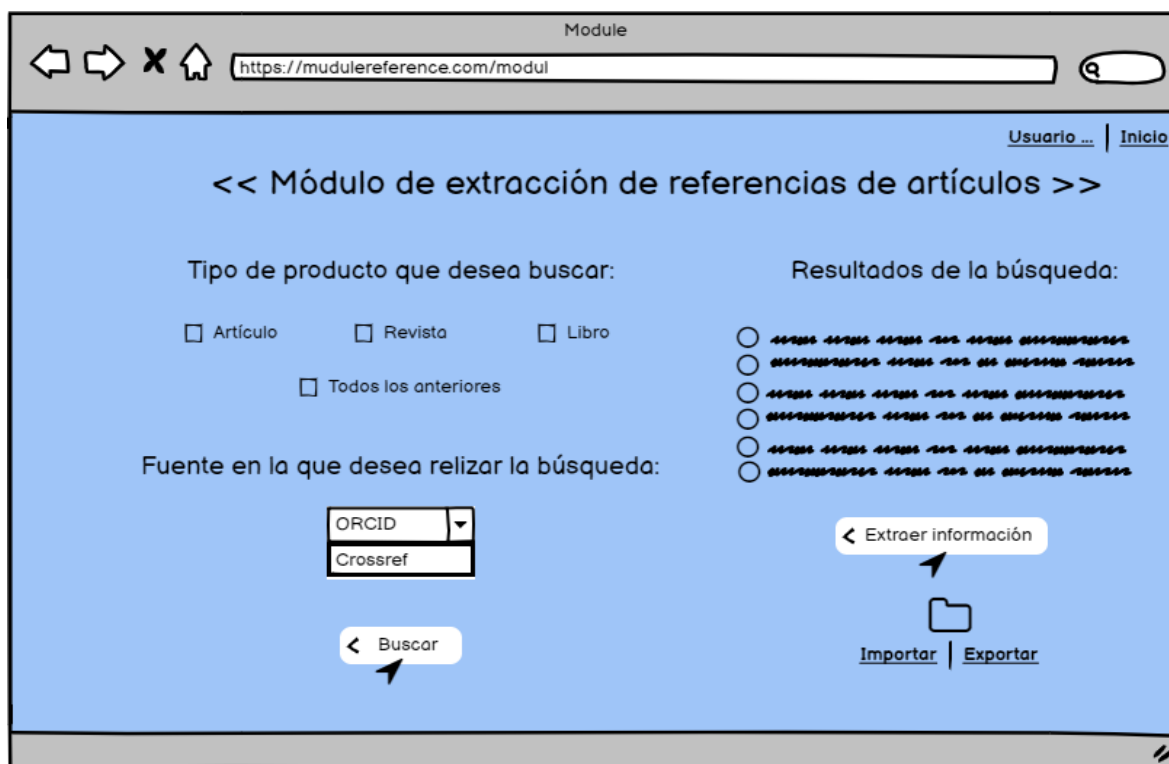


Figura 16 Mockup Interfaz de funcionamiento automatizado del módulo.

Fig. 4. Mockup de la interfaz del Módulo Web propuesto.

Las organizaciones en la elección de búsqueda ponen a disposición las siguientes Apis:

API REST Pública Crossref: La cual expone los metadatos que son depositados por los miembros en la plataforma de Crossref que son importantes para localizar publicaciones por autor, cabe mencionar que dichos datos públicos se pueden utilizar sin restricción [23].

Funcionalidades de los componentes de búsqueda:

- Trabajos utilizando el DOI.
- Miembros

- Tipos de publicaciones
- Revistas

Clasificación:

- Relevancia
- Fecha
- Hora

Pruebas de búsqueda:

Numero de DOIs de artículos que contiene la revista Crossref.

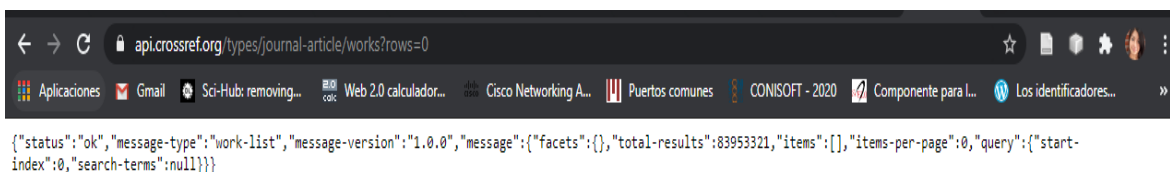


Figura 17 Búsqueda de DOI's de artículos.

Función: Búsqueda de DOIs con un prefijo determinado.

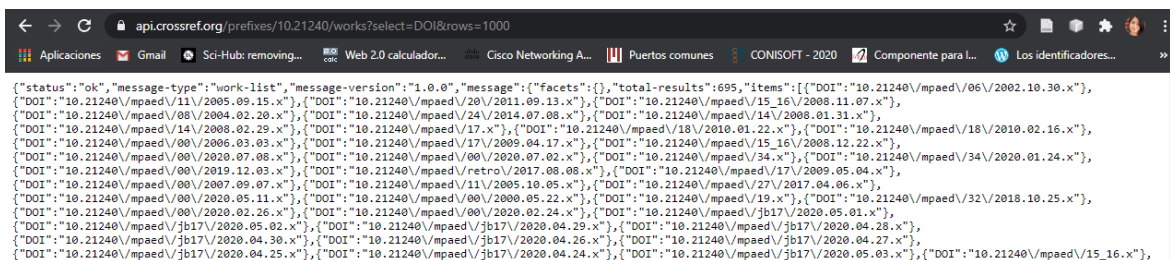


Figura 18 Búsqueda de DOI's de prefijo determinado.

Registros con la palabra sangre en los metadatos:

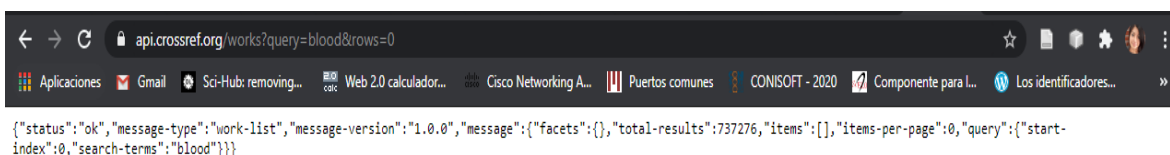
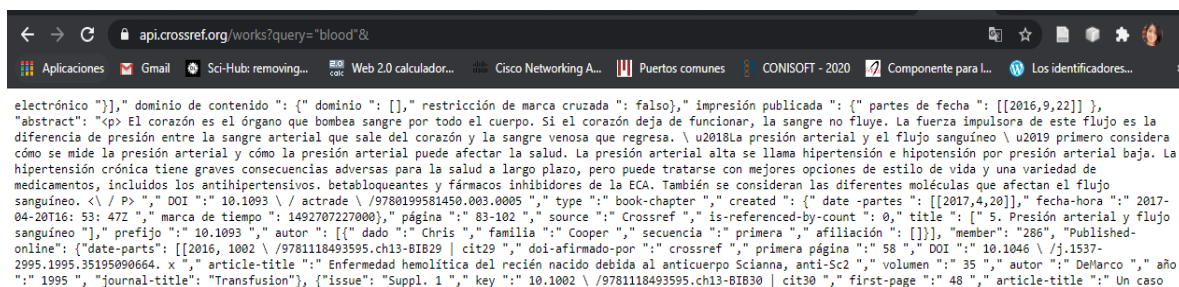


Figura 19 Búsqueda con palabras clave.

Resultados de búsqueda:



```
electrónico "}," dominio de contenido ": {" dominio ": []," restricción de marca cruzada ": falso}," impresión publicada ": {" partes de fecha ": [[2016,9,22]] }},
"abstract": "<p> El corazón es el órgano que bombea sangre por todo el cuerpo. Si el corazón deja de funcionar, la sangre no fluye. La fuerza impulsora de este flujo es la
diferencia de presión entre la sangre arterial que sale del corazón y la sangre venosa que regresa. \ u2018La presión arterial y el flujo sanguíneo \ u2019 primero considera
cómo se mide la presión arterial y cómo la presión arterial puede afectar la salud. La presión arterial alta se llama hipertensión e hipotensión por presión arterial baja. La
hipertensión crónica tiene graves consecuencias adversas para la salud a largo plazo, pero puede tratarse con mejores opciones de estilo de vida y una variedad de
medicamentos, incluidos los antihipertensivos, betabloqueantes y fármacos inhibidores de la ECA. También se consideran las diferentes moléculas que afectan el flujo
sanguíneo. <\/ P> "," DOI ": " 10.1093 \ / actrade \ / 9780199581450.003.0005 "," type ": " book-chapter "," created ": {" date -partes ": [[2017,4,20]]," fecha-hora ":" 2017-
04-20T16: 53: 47Z "," marca de tiempo ": 1492707227000}," página ": " 83-102 "," source ": " Crossref "," is-referenced-by-count ": 0," title ": [" 5. Presión arterial y flujo
sanguíneo "], " prefijo ": " 10.1093 "," autor ": [{" dado ":" Chris "," familia ":" Cooper "," secuencia ":" primera "," afiliación ":" []}], "member": "286", "Published-
online": {"date-parts": [[2016, 1002 \ / 9781118493595.ch13-BI829 | cit29 "," doi-afirmado-por ": " crossref "," primera página ": " 58 "," DOI ": " 10.1046 \ / j.1537-
2995.1995.3519590664. x "," article-title ":" Enfermedad hemolítica del recién nacido debida al anticuerpo Scianna, anti-Sc2 "," volumen ": " 35 "," autor ":" DeMarco "," año
": " 1995 "," journal-title": "Transfusion"}, {"issue": "Suppl. 1 "," key ":" 10.1002 \ / 9781118493595.ch13-BI830 | cit30 "," first-page ":" 48 "," article-title ":" Un caso
```

Figura 20 API de Crossref resultados.

API RESTful Pública ORCID: API que proporciona el acceso a la base de datos de los datos de registros ORCID ID públicos función que es elección del miembro [24].

Funcionalidades de los componentes de búsqueda:

- Obtener ORCID ID Autenticado
- Buscar
- Recuperar datos públicos

Registro en la API Sandbox

Developer tools

NOTE: This section is intended for software developers who plan to incorporate ORCID iDs into their systems. If you arrived here by mistake, you can go back to [your ORCID record](#).

You have enabled the public API

See the [Public client terms of service](#).

Name of your application (Shown to users)

Article Reference Metadata Retrieval Module

Your website URL

<http://modulereference-com.stackstaging.com/>

Description of your application

A web module as a thesis project, integrated in an institutional repository that allows to automate and integrate the retrieval of article references from non-profit academics and students.

Figura 21 Datos de registro.

Tokens de acceso asignados:

Client ID	APP-BVDAOEWDTTACBP2A
Client secret	3f64245d-0ead-4e5d-843b-850105c232ef
Authorize request	Endpoint: https://orcid.org/oauth/authorize Scope: /authenticate Response type: code Description: provides an authorization code that can be exchanged for an access token and an authenticated ORCID iD <pre>https://orcid.org/oauth/authorize?client_id=APP-BVDAOEWDTTACBP2A&response_type=code&scope=/authenticate&redirect_uri=https://pub.orcid.org/v2.0/</pre>
Token request	Endpoint: https://orcid.org/oauth/token Response type: access token and ORCID iD Description: provides an authenticated ORCID iD and an access token that can be used to read public information on the record Sample cURL <pre>curl -i -L -k -H 'Accept: application/json' --data 'client_id=APP-BVDAOEWDTTACBP2A&client_secret=3f64245d-0ead-4e5d-843b-850105c232ef&grant_type=authorization_code&redirect_uri=https://pub.orcid.org/v2.0/&code=REPLACE WITH OAUTH CODE' https://orcid.org/oauth/token</pre>
OpenID/Implicit request	Endpoint: https://orcid.org/oauth/authorize Scope: openid Response type: token Description: provides an access token that can be used to read public information on the record and an id_token using OpenID Connect and client-side only implicit OAuth. (OpenID Connect info) <pre>https://orcid.org/oauth/authorize?client_id=APP-BVDAOEWDTTACBP2A&response_type=token&scope=openid&redirect_uri=https://pub.orcid.org/v2.0/</pre>

Figura 22 Datos de acceso asignados por ORCID.

Pila de sprint

Se enlistan a continuación quince actividades para el desarrollo del módulo:

1. Investigar y analizar el estado de la práctica sobre extracción de metadatos
2. Escribir y analizar las tecnologías y estándares para dicho proceso
3. Analizar los formatos de referencia de artículos y bases de datos además conocer su estructura.
4. Escribir capítulo 1 de la de tesis.
5. Escribir capítulo 2 de la tesis.
6. Analizar los formatos de referencia de artículos y bases de datos.
7. Investigar los estándares aplicables a este desarrollo.
8. Identificar los diferentes métodos de extracción según el tipo de dato.

9. Determinar las tecnologías que se utilizarán para la implementación del producto a desarrollar.
10. Seleccionar la metodología adecuada para este tipo de desarrollo.
11. Escribir Capítulo 3 de la tesis.
12. Escribir el artículo del proyecto.
13. Validar el método implementado usando un caso de estudio
14. Realizar la estancia académica.
15. Escribir el capítulo 4 de la tesis.
16. Escribir el capítulo 5 de la tesis.
17. Entregar documento de tesis.
18. Atender correcciones de tesis.
19. Sustentar el examen de grado.

Cronograma

Se presenta la distribución de actividades en el cronograma dividido en meses y tomando en cuenta los entregables de los siguientes dos semestres:

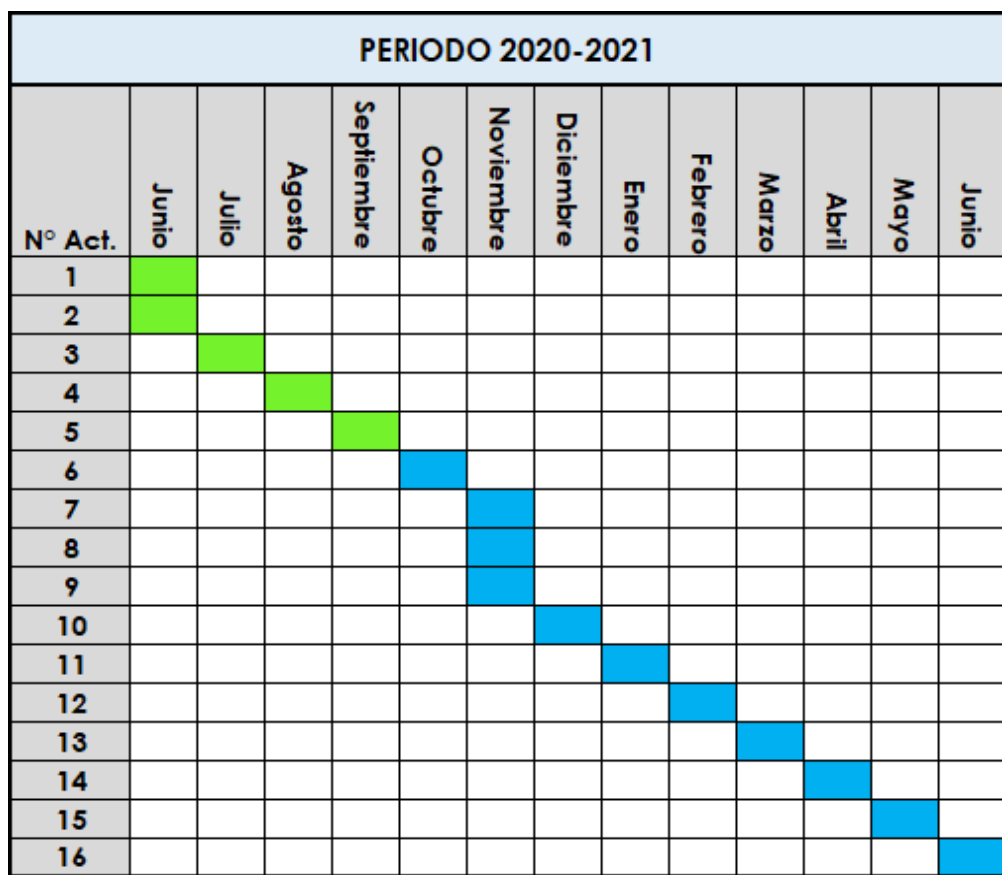


Figura 23 Cronograma de actividades 2020 - 2021.

3.4.3 Entregables

En esta última sección se muestran los entregables divididos en dos semestres.

3.4.3.1 Tercer semestre

- Diseño del módulo web para la extracción de referencias bibliográficas.
- Redacción y envío de artículo científico.
- Prototipo terminado.
- Capítulo 3 de la tesis.

3.4.3.2 Cuarto semestre

- Artículo científico aceptado y publicado.
- Módulo terminado y funcionando.
- Aplicación en un caso de estudio.
- Video y póster.
- Tesis completa.

Capítulo 4. Resultados

En este capítulo se presentan los resultados después de terminar el presente proyecto que consistió en la realización de un módulo Web para la localización referencial de los productos de investigación que son los artículos localizados desde la base de datos donde fueron publicados, utilizando los metadatos, pero principalmente el DOI.

Caso de estudio: Instituto Tecnológico de Orizaba/Tecnológico Nacional de México.

Selección del artículo: Como parte de la investigación se realizó la elección de un artículo en colaboración con esta institución que lleva entre sus referencias los siguientes datos:

Título del artículo: A study on LIWC categories for opinion mining in Spanish reviews

DOI: [10.1177/0165551514547842](https://doi.org/10.1177/0165551514547842)



Localización: Contando con el dato de identificador único del producto a localizar dentro del módulo se ingresó a la interfaz principal y se seleccionó la búsqueda mediante el DOI y se ingresó el dato:

Una vez ingresado el dato de identificador se localizó de manera referencial el artículo dentro de las bases de datos obteniendo lo siguiente:

1. DOI: "10.1177/0165551514547842"
2. ISSN: (2) ["0165-5515", "1741-6485"]
3. URL: "http://dx.doi.org/10.1177/0165551514547842"
4. abstract: "<jats:p> With the exponential growth of social media, that in each domain. </jats:p>"
5. alternative-id: ["10.1177/0165551514547842"]
6. author: (6) [{"..."}, {...}, {...}, {...}, {...}, {...}]
7. container-title: ["Journal of Information Science"]
8. content-domain: {domain: Array(1), crossmark-restriction: true}
9. created: {date-parts: Array(1), date-time: "2014-08-27T04:52:44Z", timestamp: 1409115164000}
10. deposited: {date-parts: Array(1), date-time: "2021-05-08T11:15:45Z", timestamp: 1620472545000}
11. indexed: {date-parts: Array(1), date-time: "2021-05-08T11:40:13Z", timestamp: 1620474013515}
12. is-referenced-by-count: 25
13. issn-type: (2) [{"..."}, {...}]
14. source: "Crossref"
15. subject: (2) ["Library and Information Sciences", "Information Systems"]
16. subtitle: []
17. title: ["A study on LIWC categories for opinion mining in Spanish reviews"]

Realizando la comparación de los datos obtenidos dentro de los datos obtenidos coinciden y además como dato extra nos provee del abstract o resumen del producto localizado.

<jats:p> With the exponential growth of social media, that is, blogs and social networks, organizations and individual persons are increasingly using the number of reviews of these media for decision-making about a product or service. Opinion mining detects whether the emotions of an opinion expressed by a user on Web platforms in natural language are positive or negative. This paper presents extensive experiments to study the effectiveness of the classification of Spanish opinions in five categories: highly positive, highly negative, positive, negative and neutral, using the combination of the psychological and linguistic features of LIWC (Linguistic Inquiry and Word Count). LIWC is a text analysis software that enables the extraction of different psychological and linguistic features from natural language text. For this study, two corpora have been used, one about movies and one about technological products. Furthermore, we conducted a comparative assessment of the performance of various classification techniques, J48, SMO and BayesNet, using precision, recall and F-measure metrics. The findings revealed that the positive and negative categories provide better results than the other categories. Finally, experiments on both corpora indicated that SMO produces better results than BayesNet and J48 algorithms, obtaining an F-measure of 90.4 and 87.2% in each domain. </jats:p>

El objetivo principal de esta aportación es que el módulo permita la localización referencial de los artículos producidos en este caso del Instituto Tecnológico de Orizaba para tener un control de la producción a lo largo del tiempo y las generaciones que puedan reportar a la estancia correspondiente.

Actualmente dentro de los RI se encuentran con información escasa y estadísticamente hablando no existe un reporte general por autores, edades, entre otros datos relevantes, si bien es importante recordar que se deben reportar las investigaciones que son financiadas con recursos públicos.

Conclusiones

Existen gestores de referencias y componentes que ayudan a extraer información referente a trabajos de investigación en bases de datos científicas, también existe información alojada y en operación por el Repositorio Nacional pero en algunos casos son datos desactualizados, pero en particular con el módulo se resolverán los casos de inconsistencia de datos en un nivel de metadatos esto ayudará a presentar mediante la extracción una información más limpia, sin duplicidad y datos erróneos lo cual es de gran importancia para la administración personal o grupal de las aportaciones científicas así como la confiabilidad por la utilización del identificador único de artículos.

También es importante mencionar que dicho módulo se realizó tomando en cuenta normas y estándares de calidad.

El fin de este trabajo es beneficiar no solo a personas que participen en algún tipo de investigación como lo hacen los estudiantes al momento de concluir un grado de estudio, sino a grupos que se especializan y encargan de estudios de investigación de mayor nivel, los cuales podrán extraer y acumular sus trabajos, para ello se utilizarán las interfaces de aplicaciones para acceder desde un solo entorno a las bases de datos que hagan uso de este tipo de información, es necesario recordar que debido a las estipulaciones y el acceso abierto el unico fin es que no exista duplicidad y se promueva la innovación en nuestro país.

Bibliografía

- [1] «Gestores de referencias – PoliScience». <https://poliscience.blogs.upv.es/investigadores-2/mis-citas/gestores-de-citas/> (accedido ago. 11, 2020).
- [2] «Mendeley - Software de gestión de referencias y red de investigadores». https://www.mendeley.com/?interaction_required=true (accedido ago. 11, 2020).
- [3] «Zotero | Tu asistente de investigación personal». <https://www.zotero.org/> (accedido ago. 11, 2020).
- [4] «EndNote | Clarivate Analytics». <https://endnote.com/> (accedido ago. 11, 2020).
- [5] «Zona de Lectura | Elsevier». <https://www.elsevier.es/es> (accedido ago. 12, 2020).
- [6] «¿Qué es un Plugin y para que sirve? - Neo Wiki», *NeoAttack*. <https://neoattack.com/neowiki/plugin/> (accedido nov. 05, 2020).
- [7] «Agencias de registro de DOI». https://www.doi.org/registration_agencies.html (accedido jun. 14, 2020).
- [8] «ORCID». <https://orcid.org/> (accedido ago. 12, 2020).
- [9] A. T. Toukourmidis, «Los identificadores de autor, revista y artículo: ORCID, ISSN y DOI», *Revista Comunicar. Escuela de Autores*, oct. 17, 2016. <https://comunicarautores.com/2016/10/17/los-identificadores-de-autor-revista-y-articulo-orcid-issn-y-doi/> (accedido mar. 04, 2020).
- [10] R. PowerData, «¿Qué son los metadatos y cuál es su utilidad?». <https://blog.powerdata.es/el-valor-de-la-gestion-de-datos/que-son-los-metadatos-y-cual-es-su-utilidad> (accedido nov. 05, 2020).
- [11] L. J. Wilkinson, «You are Crossref», *Crossref*. <https://www.crossref.org/> (accedido ago. 12, 2020).
- [12] «TPM | Tutorial de Programacion Multiplataforma». <https://itslr.edu.mx/archivos2013/TPM/temas/s3u6.html> (accedido ene. 11, 2021).
- [13] M. Klein y H. Van de Sompel, «Discovering Scholarly Orphans Using ORCID», en *2017 ACM/IEEE Joint Conference on Digital Libraries (JCDL)*, Toronto, ON, Canada, jun. 2017, pp. 1-10, doi: 10.1109/JCDL.2017.7991573.
- [14] H. Holzmann y M. Runnwerth, «Micro Archives as Rich Digital Object Representations», en *Proceedings of the 10th ACM Conference on Web Science*, Amsterdam Netherlands, may 2018, pp. 353-357, doi: 10.1145/3201064.3201110.
- [15] «Significado de URL (Qué es, Concepto y Definición) - Significados». <https://www.significados.com/url/> (accedido dic. 03, 2020).
- [16] D. Bangert y M. Frances, «PIDs to Support Discovery and Citation: Persistent Identifier Service Design and Delivery at UNSW Library», en *2017 ACM/IEEE Joint Conference on Digital Libraries (JCDL)*, Toronto, ON, Canada, jun. 2017, pp. 1-2, doi: 10.1109/JCDL.2017.7991610.
- [17] D. Yang y A. N. Zhang, «Performing literature review using text mining, Part III: Summarizing articles using TextRank», en *2018 IEEE International Conference on Big Data (Big Data)*, Seattle, WA, USA, dic. 2018, pp. 3186-3190, doi: 10.1109/BigData.2018.8622408.
- [18] «Run JavaScript Everywhere.» <https://nodejs.dev/> (accedido ene. 12, 2021).

- [19] «NetBeans IDE 8.2 Release Information». <https://netbeans.org/community/releases/82/> (accedido ene. 12, 2021).
- [20] «MariaDB Foundation - MariaDB.org». <https://mariadb.org/> (accedido ene. 12, 2021).
- [21] «What is Scrum?», *Scrum.org*. <https://www.scrum.org/resources/what-is-scrum> (accedido ago. 11, 2020).
- [22] «Qué es una API». <https://hipertextual.com/archivo/2014/05/que-es-api/> (accedido ago. 12, 2020).
- [23] «REST API - Crossref». <https://www.crossref.org/education/retrieve-metadata/rest-api/> (accedido nov. 06, 2020).
- [24] L. Krznarich, «ORCID API», oct. 23, 2014. <https://orcid.org/organizations/integrators/API> (accedido nov. 06, 2020).